TRINUCLEOTIDE REPEAT LENGTH DISTRIBUTION AND MITOCHONDRIAL DNA HAPLOGROUP IN SUB-ETHNIC GROUP OF NEWAR POPULATION OF NEPAL

M.Sc. Thesis
(2017)

Submitted to:

CENTRAL DEPARTMENT OF BIOTECHNOLOGY
TRIBHUVAN UNIVERSITY
Kirtipur, Kathmandu, Nepal

By
Medha K.C.

Registration No: 5-2-282-77-2009

Supervisors

Prof. Dr. Tilak R. Shrestha
Central Department of Biotechnology
Tribhuvan University, Kirtipur

Dr. Mohammed Faruq
Scientist, MBBS, Ph.D
CSIR-Institute of Genomics and Integrative Biology, New Delhi, India

Acknowledgement

I am greatly indebted to my supervisor Prof. Dr. Tilak R. Shrestha, Central Department of Biotechnology, Tribhuvan University for providing me opportunity to work at CSIR-IGIB, New Delhi, his expert balance of encouragement and constructive criticism and overall supervision of thesis writing. The research programme is a part of his long term vision of Human Variome Project International (HVPI) and collaboration between IGIB, New Delhi and CDBT-TU to work on vast array of Nepalese genetic diseases in future. My sincere thanks to Dr. Mohammed Faruq, Scientist, and CSIR-IGIB for allowing me to conduct my 6 month thesis experiment in his lab. Also, I am very thankful to his continuous expert inputs, support and encouragement throughout.

I am equally indebted to my home institution, Central Department of Biotechnology, Tribhuvan University and would also express my gratitude to Prof. Dr. Krishna Das Manandhar (Head of Department) and Prof. Dr. Rajani Malla (former Head of department), Central Department of Biotechnology, Tribhuvan University for providing official process to undertake present research study. I am equally grateful to Raman Krishna Maharjan, Sandesh Maharjan, Archana Maharjan, Nutan Thakur, Binod Neupane for arranging the blood collection of Newar community (Maharjan) and providing help to conduct my research. I am indebted to all respondents during the blood collection. I would also like to acknowledge my senior Nagendra Awasthi for his valuable time for completion of my thesis work.

My special thanks to Mohan Bahadur Shrestha, Milan Mainali, Gauri Thapa, Sujan Lamichhane, Mitesh Shrestha, Mukesh Thapa and all my classmates, seniors, juniors and my lab mates at CSIR-IGIB Renu Kumari, Dr. Aradhna Mathur, Varun Suroliya for their continuous help and support during my thesis work.

Furthermore, I would sincerely acknowledge the contribution of my parents and bestow my heartiest appreciation for their encouragement. Without their greatest support I wouldn't be able to conduct my thesis work in CSIR-IGIB, New Delhi.

LIST OF ABBREVIATIONS

APS	Ammonium persulfate
bp	Base pair
ddNTP	di-deoxy Nucleotide triphosphate
DM1	Myotonic dystrophy-1
DM2	Myotonic dystrophy-2
DNA	Deoxyribo nucleic acid
dNTP	deoxy Nucleotide triphosphate
DRPLA	Dentatorubral-pallidoluysian atrophy
EDTA	Ethylene diamine tetra acetic acid
FAM	Fluorescein amidite
FRAX-E	Fragile X mental retardation
FRDA	Friedreich's ataxia
FXS	Fragile X syndrome
FXTAS	Fragile X- associated tremor and ataxia syndrome
HD	Huntington's disease
HDL2	Huntington's disease-like 2
HEX	Hexachloro-Fluorescein
HVR I	Hypervariable region I
HVR II	Hypervariable region II
LN	Large Normal
MQ	MilliQ
MJD	Machado-Joseph Disease
mtDNA	Mitochondrial DNA
NaCl	Sodium Chloride
NaOAc	Sodium Acetate
NaOH	Sodium Hydroxide
NLB	Nuclei Lysis Buffer
np	Nucleotide position
OPMD	Oculopharyngeal muscular dystrophy
OXPHOS	Oxidative phosphorylation system

PCR	Polymerase Chain Reaction
PEG	Poly Ethylene Glycol
RBC	Red Blood Cell
RFLP	Restriction Fragment Length Polymorphism
RLB	RBC Lysis Buffer
SCA	Spinocerebellar ataxia
SDS	Sodium dodecyl sulphate
SMBA	Spino-muscular bulbar atrophy
STR	Short Tandem Repeats
TRE	Trinucleotide repeat expansion
VNTRs	Variable Numbers of Tandem Repeats
YBP	Years before present

Contents

List of Figures:

List of Tables:

ABSTRACT

Tandem nucleotide repeats are repetitive DNA in which two or more contiguous, approximate copies of a pattern of nucleotide occur in a DNA sequence. Trinucleotide repeats are a form of tandem repeat which are caused by an expansion of a segment of DNA that contains a repeat of 3 nucleotides (Triplet repeat). Variable number of triplet repeats are constituted in a healthy individual but there is a threshold beyond which a high number of repeats causes disease. This threshold varies in different disorders. Trinucleotide repeat expansions are highly polymorphic and sometimes called dynamic or unstable mutation because the number of repeats increases as the gene passes from parents to offspring. Expansion of disease arises from existence of large normal of normal repetitions (Large Normal alleles). Frequency of large normal alleles' estimation is the indirect measure of the prevalence of the disease. To investigate the normal allele range of trinucleotide repeat disorder (SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, DM1, DRPLA, FXTAS and HD) blood samples from 55 healthy unrelated individual belonging to Newar sub-ethnic group, the Maharjan of Nepal were examined. PCR products were subjected to capillary electrophoresis (ABI 3130xl Genetic Analyser) for fragment analysis. In the studied population the normal range of 10 different loci were SCA1, 22 to 36 (CAG)n ; SCA2, 19 to 26 (CAG)n; SCA3, 14 to 36 (CAG)n ; SCA7, 6 to 14 (CAG)n; SCA8, 11 to 32 (CTG)n; SCA12, 9 to 18 (CAG)n; HD, 12 to 28 (CAG)n; FXTAS , 18 to 34 (CGG)n; DRPLA, 7 to 22 (CAG)n; DM1, 1 to 28 (CTG)n .

Variability of human mitochondrial DNA has provided valuable data about the genetic history of human. Analysis of the frequency, variation and distribution of mitochondrial DNA haplogroup have been used to evaluate genetic structure of various population residing in different geographical region. Second part of this dissertation concentrates on the mitochondrial DNA haplogroup determination of Newar sub-ethnic group, the Maharjan of Nepal. Mitochondrial D-loop is the non-coding region which has two hypervariable regions that exhibit very high mutation rates, thus, can distinguish recently diverged population. In this study, D-loop region was sequenced and Individual lineages were constructed on the basis of mtDNA mutations. 5 major haplogroups with different frequencies were observed in mtDNA haplogroup viz. M (58.2%), N (18.2%), H (3.6%), R (12.7%), U (7.3%). Our results indicated presence of South Asian specific haplogroup M, including M3, M5, Z, G etc. in high frequency along with branches R (R9), U (U7, U8). These results indicated Maharjan gene pool was found to harbour almost 57.56% of South Asian, 19.99% of East Asian, 18.85% of Western Eurasian and 3.6% of Central Asia specific gene pool.

Key words: Trinucleotide repeat disorder, Haplogroup, Gene pool, D-Loop

CHAPTER 1

INTRODUCTION

1.1 Background

The analysis of genetic diversity and relatedness between or within different species, populations and individuals is a central task for many disciplines of biological science. The human genome exhibits a large amount of diversity, within and between populations and individuals. This variation manifests itself in a number of ways, for instance as insertions or deletions, variable numbers of tandem repeats (VNTRs) or single nucleotide polymorphisms (SNPs) (Rao et al., 2010). Variation can simply be defined as differences and deviations from the reference. Understanding of these variations in molecular levels has now become a challenging task in human genetic research and several approaches have been made to solve these problems.

In 1985, it was discovered that DNA contains highly polymorphic regions of repeated sequences that can be visualized by Jeffrey probes (Kloosterman, 2003). Restriction Fragment Length Polymorphism (RFLP) was used to explore human genome. Minisatellites pattern revealed by Jeffreys probes on unknown loci throughout the genome and were called Multi-Locus Probe (MLPs). This was applied for human identification in forensic and paternity determination. Later in 1988, the Single Locus Probes (SLPs) were introduced. Combinations of SLPs were used to increase discriminating power and by using these, population databases could be generated. After 1993, New PCR based technique was introduced. This was based on Short Tandem Repeats (STRs) of 2-6 bp. Contrary to RFLP analysis that requires microgram amounts of intact DNA, PCR allows the amplification and detection of sub-nanogram amounts of lower molecular weight DNA in hours rather than days (Kloosterman, 2003).

Major demographic events like migration, population bottlenecks and population expansion leave genetic imprints where gene frequency of the genome is altered (Kivisild, 1998). These imprints are passed onto successive generations thus preserving population history within the population. During 1980's and 1990's when the human genome sequencing had not been completed mtDNA become the focus of evolutionary studies. Maternal inheritance, rapid mutation rate, high copy number per cell and the lack of recombination were the features of mtDNA that provide valuable data about genetic history of human. Most studies of human evolution is based on mtDNA sequencing of control region that constitute less than 7% of the mitochondrial genome. The data generated from Non-coding region (control region) of mtDNA are combined into the shape of phylogenetic tree. The branches of phylogenetic tree are assigned alphabetical labels known as mtDNA haplogroups. The nomenclature of mtDNA haplogroups was introduced in the mid-1990s with A-G labels assigned to variation

observed in Asian and American lineages, H-K to Europe whereas only a single letter L was assigned to describe the highest level of variation observed in Africa (Kivisild, 2015).

1.2 Tandem Nucleotide Repeats loci in Human Genome

Eukaryotic genomes are full of repeated DNA sequences called satellite DNA. These repeated stretches of DNA sequences come in variable types and sizes, consisting of a core repeat unit that is tandemly repeated. Such polymorphic sequences have proved useful as markers in linkage studies. Expansion of one group of simple repeats, trinucleotide or triplet repeats, is now known to cause about 20 inherited disorders which includes DRPLA (Dentatorubropallidoluysian Atrophy), Huntington's Disease (HD), SBMA (Spinobulbar Muscular Atrophy), Spinocerebellar Ataxias, Fragile X Syndrome, Fragile XE Mental Retardation, Friedreich's Ataxia, Myotonic Dystrophy (Paulson et al., 1996). Tandem repeats in nucleotide sequences are two or more adjacent and approximate copies of a sequence of nucleotides. They are relatively common and up to 10% can exist within protein coding gene. Different types of repetitive elements can account for up to 50% of the genome. The presence of tandem repeats and variations within these repeats have been associated with a large number of diseases and phenotypic outcomes (Lander et al., 2001).

1.2.1 Classification of Tandem Nucleotide Repeats:

Tandem repeats often have been classified as followings based on their lengths:

Satellites: These are the tandemly repeated DNA sequences located in pericentromeric and telomeric regions of the heterochromatin which can be upto hundreds of bp per repeat (Rich et al., 2014)

Minisatellites: Minisatellites are the tract of repetitive DNA in which DNA motifs ranging in length from 10- 60 base pairs are typically repeated 5 – 50 times. They occur at more than 1000 locations in human genome and are notable for high mutation rate and high diversity in the population. These are precisely associated with telomere in terms of their location (Butler et al., 2005).

Microsatellite: These are the tandemly repetitive DNA sequence characterized by their short sequence repeat length of <10 bp. They are also referred as Simple Sequence Repeats (SSRs) or Short Tandem Repeats (STRs) (Butler et al., 2005).

Table 1: Types of repeated DNA sequences in Human Genome (Rich et al., 2014)

Type of DNA	Description	Length
Macrosatellite	type of satellite that are often specific for only one or two chromosomes	>1000 bp
Satellite	DNA Regions with long stretches of repeated DNA sequences, mostly found along the centromeres	~100 bp
Minisatellite	type of satellite DNA consisting of medium length repeat units	10-60 bp
Microsatellite	type of satellite DNA consisting of smaller repeat units the so-called Simple Tandem Repeat, or Short Tandem Repeats(STR)	<10 bp

1.2.2 Trinucleotide Repeats:

Occurrence of simple sequence repeats is common in the human genome. Microsatellites or tandem repeats are polymorphic sequences which underlie an entirely new class of human mutations. Trinucleotide repeat is the stretch of three nucleotides repeated multiple times in a DNA sequence (Paulson & Fischbeck, 1996). It has been estimated that as many as 1 in 20 human proteins have tandem repeat polymorphisms & approximately one fifth involve tandem repeat units that are not multiples of three. The expansion of unstable trinucleotide repeats can cause neurological disorders & now accounts for at least 16 neurological disorders. The repeat can either located in an exons or outside the open reading frame but both can have significant impact (Orr & Zoghbi, 2007).

Table 2: Molecular characteristics of Trinucleotide Expansion in humans (McMurray, 2010)

DISEASE	SEQUENCE	LOCATION	NORMAL REPEAT NUMBER	PRE-MUTATION	PATHOGENIC
Disease with non-coding TNRs					
DM1	CTG	DMPK (3'UTR)	5-37	37-50	>50
DM2	CCTG	*CNBP* (INTRON 1)	<30	31-74	>75
FRAX-E	GCC	*AFF2* (5'UTR)	4-39	40-200	>200
FRDA	GAA	*FXN* (INTRON 1)	5-30	31-100	70-1000
FXS	CGG	*FMR1* (5'UTR)	6-50	55-200	200-4000
HDL2	CTG	*JPH3* (EXON 2A)	6-27	29-35	36-57
SCA8	CTG	*ATXN8OS* (3' UTR)	15-34	34-89	89-250
SCA10	ATTCT	*ATXN10* (INTRON9)	10–29	29–400	400–4,500
SCA12	CAG	*PPP2R2B* (5' UTR)	7-28	28-66	66-78
Diseases with Coding TNRs					
DRPLA	CAG	*ATN1* (EXON 5)	6-35	35-48	49-88
HD	CAG	*HTT* (EXON 1)	6-29	29-37	38-180
OPMD	GCN	*PABPN1* (EXON1)	10	12-17	>11
SCA1	CAG	*ATXN1* (EXON 8)	6-39	40	41-83
SCA2	CAG	*ATXN2* (EXON 1)	<31	31-32	32-200
SCA3	CAG	*ATXN3* (EXON 8)	12-40	41-85	52-86
SCA6	CAG	*CACNA1A* (EXON 47)	<18	19	20-33
SCA7	CAG	*ATXN7* (EXON 3)	4-17	28-33	>36
SCA17	CAG	*TBP* (EXON 3)	25-42	43-48	45-66
SMBA	CAG	*AR* (EXON 1)	13-31	32-39	40

1.3 Human Genome Diversity:

Human genetic diversity has important implications for human evolution, forensics, and the distribution of genetic diseases in populations. mtDNA has been a widely used tool in human evolutionary and population genetic studies over the past three decades. Its maternal inheritance and lack of recombination have offered the opportunity to explore genealogical relationships among individuals and to study the frequency differences of

matrilineal clades among human populations at continental and regional scales (Kivisild, 2015)

1.3.1 Mitochondrial Genetics:

Only extra nuclear source of DNA in animal cells is mitochondria. mtDNA is a circular, double-stranded, 16569 bp molecule of DNA which encodes 37 genes, including 13 essential polypeptides for the OXPHOS system, two ribosomal RNAs (12S and 16S) and 22 tRNAs. Unique characteristics that distinguish it from nuclear genome are; it is maternally inherited and there are several hundred to several thousands of copies within a single cell (Taylor et al., 2005).

The mtDNA is replicated and transcribed by using an origin and a promoter for each of the two DNA strands, the G-rich heavy (H) strand and the C-rich light (L) strand. The H- and L-strand origins (O_H and O_L) are separated by 2/3 of the molecule, but the H- and L-strand promoters (P_H and P_L) are located adjacent to O_H in the approximately 1000-np noncoding control region, which also encompasses the triple-stranded D-loop (Wallace, 1994).

The displacement loop (D-loop) is the only major non-coding region of mtDNA and is formed by the displacement of the two genomic strands by a third DNA strand. The D-loop is a 1.1 kb region containing the important regulatory elements for mtDNA transcription and replication. The promoters and leading-strand origin are located near the 5'-boundary of the D-loop structure. This entire locus is commonly referred as the D-loop regulatory region or mtDNA control region (Shadel & Clayton, 1997). D-loop structure is created by nascent short heavy (H) strand which displaces the parental H strand. Because of its reduced size and improved knowledge in the replication and transcription mechanisms and the availability of sequences from many species, it represents a good model for studying the evolution (Sbisà et al., 1997).

Mitochondrial DNA is inherited unilaterally via the maternal line and is present in nearly all cells. Within cells, mtDNA is present in high copy numbers and most of these copies are identical to one another. Cells which require a greater ATP yield, such as muscle or nerve cells, will contain a greater number of mitochondria and therefore more mtDNAs than those which have a much lower ATP demand (Taylor et al., 2005).

1.3.2 Haplotype:

A haplotype is a group of genes within an organism that was inherited together from a single parent. In addition, the term "haplotype" can also refer to the inheritance of a cluster of single nucleotide polymorphisms (SNPs), which are variations at single positions in the DNA sequence among individuals. A haplotype can refer to a combination of alleles or to a set of single nucleotide polymorphisms (SNPs) found on the same chromosome (Gibbs et al., 2003).

1.3.3 Haplogroup:

Haplogroup is a group of similar haplotypes that share a common ancestry, possessing the same single nucleotide polymorphism (SNP) mutation. Because a haplogroup consists of similar haplotypes, it is possible to predict a haplogroup from haplotypes. In human genetics, the most commonly studied haplogroups are Y-chromosome(Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations (Gibbs et al., 2003).

1.3.4 Mitochondrial Haplogroups:

Difference in the mitochondrial DNA helps to understand the evolution of female lineage and to trace the matrilineal inheritance of modern human back to human origins in Africa. Haplogroup are assigned by letters from A to Z. In human mitochondrial genetics, haplogroup L is the root of the human mtDNA phylogenetic tree. The major sub-clade of haplogroup L is L3 which represents the most common maternal lineage of all people outside of Africa. L3 is subdivided into haplogroups M and N from which vast majority of non-africans are descended. Macrohaplogroup M is found in higher frequency in Asia with a frequency range of 60%-80%. Also called as sibling haplogroup of M, haplogroup N is also descended of haplogroup L3. Haplogroup N is the ancestral haplogroup to almost all European and Oceanian haplogroup (Mishmar D et al., 2003).

1.3.5 Population under Study:

Nepal is divided into three regions i.e. Himalayan region- located along southern slopes of the Himalayas, Terai region- predominantly Hindu and consists of low altitude fertile plains and between two extremies lie the intermediates hills and valleys where a majority of the Nepalese population resides. Nepalese society is the one of the world's most diverse and complex; consisting about 123 different ethnic group (Nepal census report, 2011). According to historical records, it was thought that by the seventh century the Newar tribes controlled Nepal, which then consisted only of the Kathmandu Valley, located in the east central hills of present- day Nepal. Newars are spread across the country but majority are still concentrated in the Kathmandu valley. A Tibeto- Burman language, Nepalbhasa is spoken by Newar.

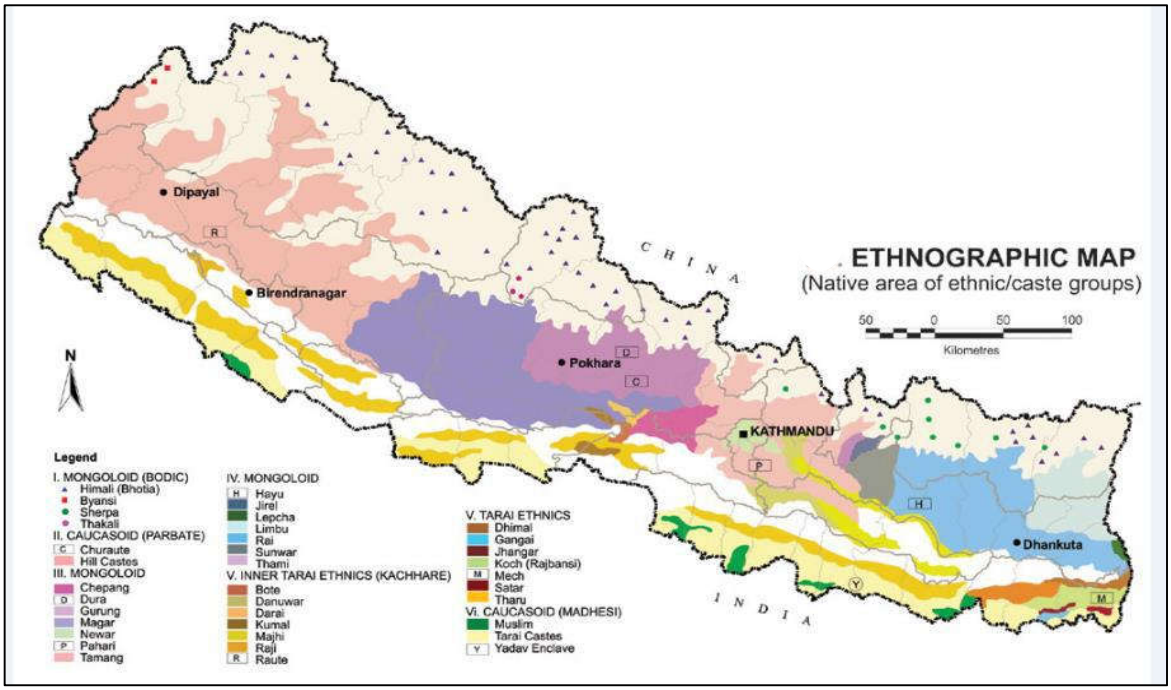


Fig. 1: Map of Nepal describing different ethnicities (www.nepalitimes.com).

1.4 Hypothesis:

Tandem repeat expansion has a characteristics feature of dynamic mutation where the size of an initial expansion determines the rate of further expansion. Trinucleotide repeats are common form of tandem repeats expansion causing various neuromuscular and neurodegenerative diseases. Trinucleotide disorders are known to harbour alleles of large size in the normal range (large normal alleles) that are unstable and do undergo expansion to reach an intermediate range from which further expansion to the disease range takes place. These alleles have been considered as reservoirs for the generation of new expanded alleles. Distribution of normal range of repeats varies from one population to another. A study on the percentage of large normal alleles in any population would be an indirect measure of the prevalence of the disease in that population. On the other hand, human genetic diversity has an important implication for distribution of genetic disease in a population. In this study, genetic linkage study of Maharjan population can reveal an insight of their genetic diversity and moreover, correlation between genetic analysis of trinucleotide expansion and genetic diversity can be done.

1.5 Objective:

1.5.1 General Objective

- To assess the length distribution variations in disease associated trinucleotide repeat loci & to determine mitochondrial haplogroup among Newar sub-ethnic group, the Maharjan population of Nepal.

1.5.2 Specific Objectives

1. To amplify specific disease associated trinucleotide repeat loci and perform fragment analysis of amplified product.

2. To determine the normal range of repeats of various trinucleotide repeat associated disease.

3. To assign the mtDNA haplogroup according to mitochondrial DNA variations and analyse genetic diversity.

`

CHAPTER 2

REVIEW OF LITERATURE

2.1 Trinucleotide Repeat Expansion:

Human genome consists of 3.2 billion DNA base pairs with many recurring motifs of repetitive DNA. These include microsatellite repeats, repetitive DNA sequences in the telomeres, centromeres, and heterogeneous DNA regions of the chromosome (Lutz, 2007). Microsatellites constitute 3% of the human genome, but the triplet repeat has taken on special significance due to its highly unstable nature. Classical (Mendelian) genetics is based on the principle that mutations are stably transmitted between generations. However, a different type of inheritance was described for a human neurological disorder named myotonic dystrophy- characterized by increased expressivity. A similar hereditary pattern was later observed for other neurological disease like Huntington's disease, spinal and bulbar muscular atrophy, and several ataxias (Mirkin, 2007).

Alterations in the lengths of repetitive DNA over evolutionary time scales create diversity in the species. Dynamic & unstable transmission of simple repetitive elements in DNA is a new type of mutation, which has changed the face of genetics. The mutation, referred to as "Trinucleotide repeat (TNR) expansion," occurs when the number of triplets present in a mutated gene is greater than the number found in a normal gene (Hedge & Saraph, 2011). Mammals have developed systems for resisting rapid changes that could be deleterious. These simple triplets beyond a critical threshold length override genomic safeguards and expand during most parent–child transmissions and during the lifetime of an organism. Over the years, diseases associated with insertion and deletions of microsatellite sequences have grown, and the tracts in these disease's genes do not exclusively comprise triplet units. However, the consequences of triplet instability on human health are profound. Unstable simple repeats at or near genes leads to the mutation underlying dozens of severe neuromuscular and neurodegenerative disorders (La Spada & Taylor, 2010).

2.1.1 Instability of Trinucleotide Repeats:

The list of neurological disorders caused by unstable repeats has increased not only due to trinucleotide repeats but also due to tetranucleotide and pentanucleotide repeats. Instability of the size of DNA triplet repeats leads to a gradual expansion and molecular pathological effects that may cause diseases. Unstable expansion of simple DNA repeats underlies about 20 severe neuromuscular and neurodegenerative disorders. Triplet repeats may undergo further expansion or contraction depending on individual disease process, triplet repeat type & parent of origin of triplet repeat. Expansion of triplet repeat size may give rise to pre-mutations allele that has increased susceptibility to mutation in a subsequent transmission (Mirkin, 2006).

Anita Harding was the first to identify the correlation between trinucleotide repeat expansion and diseases causing neurological dysfunction. In 1991, Fragile X Syndrome was shown to be caused by an expansion of a repeated triplet DNA sequence (CGG). Discovery of CGG expansion in the 5' untranslated region of the Fr*agile X Mental Retardation 1 (FMR1)* in Fragile X Syndrome (FXS), CAG expansion in the coding sequence of X-linked Spinal and Bulbar Muscular Atrophy (SMBA), CTG expansion in the 3' untranslated region of *Myotonic Dystrophy Protein Kinase (DMPK)* in Myotonic Dystrophy Type 1 (DM1), and CAG expansion in the exon 1 coding sequence of Huntington's disease (HD) provided evidence that expansion of triplet tracts was the underlying mutation in disease (Budworth & Mcmurray, 2013).

DNA tandem repeats may be within the coding region or outside the open reading frame (Fig. 2). Among various neurodegenerative trinucleotide repeat disorder polyglutamine repeat disorder is the common one. Polyglutamine repeat disorders are expanded polyglutamine (polyQ) tract which results in formation of protein aggregations within the cell.

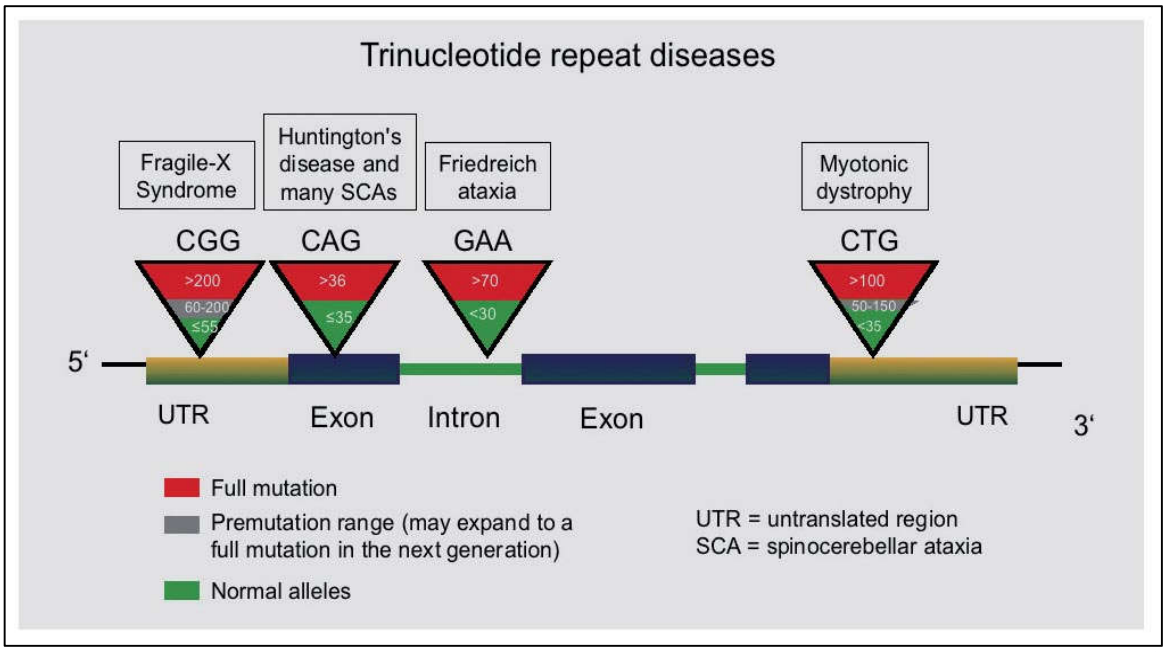


Fig. 2: Location of triplets causing different trinucleotide disorder (Source: http://www.researchgate.net/)

2.1.2 CAG Repeats:

Majority of patients with common feature of polyglutamine disorders are due the expansion of CAG repeats beyond a threshold of 35 - 40. Among various CAG repeats in complete human genome only few could be putative candidate for polyglutamine disorders. CAG loci implicated in polyglutamine disorders are polymorphic, highly conserved and harbours interruptions in repeat tract that are lost in expanded allleles. These are expressed in brain and are prone to aggregation (Pandey et al., 2004).

2.1.2.1 Open Reading Frame Expansions:

At least 10 diseases have been identified with triplet repeat expansion within exon. The expansion is an increase in the length of a polyglutamine (PolyQ) tract in the encoded protein which results in the death of vulnerable neurons in the brain. PolyQ is thought to cause conformational changes that confer toxic properties to the protein by appearance of polyQ- containing inclusions (Usdin, 2008). Failure to properly degrade misfolded polyQ proteins either via autophagy or ubiquitin-proteasome pathways contribute to polyQ toxicity (Orr & Zohgbi, 2007).

2.1.2.2 Non Coding Expansions:

There are repeat expansion diseases that involve a repeat in non-coding region of the gene. The relationship between repeat expansion and disease pathology is not just due to the change in the properties of the protein product of affected gene. The non-coding repeats affect cell and chromatin structure as well as transcription, splicing, and translation in a variety of different ways. Non-coding expansion is majorly responsible for down regulation of respective gene. The mechanism responsible for perturbed gene expression in these diseases remain unclear (Usdin, 2008).

2.1.3 Causes for Repeat Expansion:

Tandem repeats are often highly polymorphic. Mutation rate of tandem repeats can range from 10^{-2} to 10^{-6} events per locus per generation. However, mechanism by which polymorphism arise varies. They can arise from events such as gene conversion during recombination or unequal crossover, replication slippage, or double-strand break repair (recombination-based mechanism) (Dúshláine, 2006).

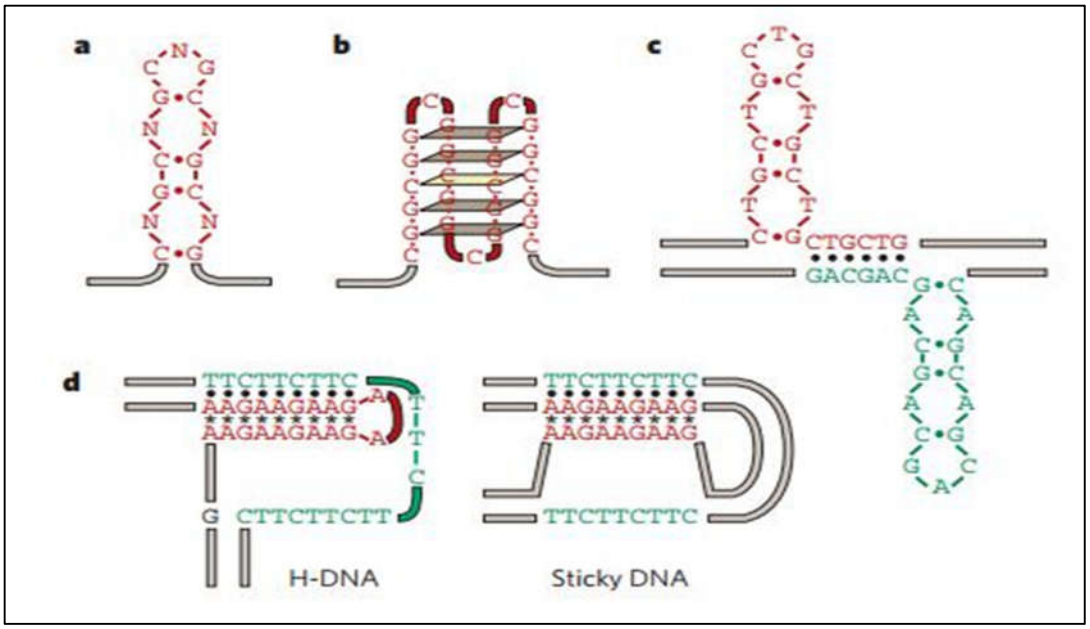


Fig 3: Unusual DNA structures formed by expandable repeats a) imperfect hairpin formed by (CNG)n repeats **b**) a quadruplex like structure formed by (CGG)n repeats **c**) a slipped stranded structure **d**) H-DNA and sticky DNA formed by (GAA)n. (TTC)n repeat (Source: Mirkin, 2007).

2.1.3.1 Polymerase Slippage during Replication:

The first molecular model of how repeat expansions occur was based on DNA polymerase slippage during replication. Looping out one or several repeats in the newly synthesized DNA strand should covert the loop into expansions after a second replication. Single-stranded (CNG)n repeats form hairpin like structures that consists both Watson- Crick base pairs and mismatched base pairs (Mirkin, 2006).

The denaturation and renaturation of double-stranded DNA fragments that contain expandable repeats promote the formation of the 'slipped-stranded' DNA conformation. During replication when DNA polymerase encounters a direct repeat, the polymerase complex suspends replication. In this case, an out-of-register realignment of the complementary repetitive strands gives rise to 'slipouts' that are folded into hairpin-like structures. DNA polymerase reassembles its position on the template strand and resumes normal replication, but during the course of reassembling, the polymerase complex backtracks and repeats the insertion of deoxy ribonucleotides that were previously added leading to expansion of repeats (Pearson et al., 2002). However, this hypothesis alone could not adequately explain several characteristics of repeat expansions since; strand slippage usually causes limited repeat length polymorphism, rather than large-scale expansion (Pearson et al., 2002).

2.3.1.2 Base Excision Repair:

Several DNA repair mechanisms, like mismatch repair and transcription-coupled repair (also known as nucleotide excision repair), have been proposed to be involved in TNR expansion. In recent studies, repair of DNA lesions produced by oxidative stress has been considered to be key factor in TNR expansion. Oxidative stress can be caused by any chemically reactive oxygen molecule that includes superoxide anion, hydroxyl radicals and peroxide. Oxidised base damage is a type of genomic DNA lesion that commonly results from the action of reactive oxygen species. These lesions, if not repaired, can lead to mutations, repeat instability and abnormal gene transcription which ultimately can cause adverse effects. To overcome the adverse effect mammalian cells maintains DNA repair mechanism, base excision repair to remove oxidative lesions. 8-oxoguanine (8-oxoG) is a frequently produced form of oxidative base in mammals. The repair of 8-oxoG lesions is initiated by 8-oxoguanine DNA glycosylase (OGG1), which removes the 8-oxoG lesion. OGG1 releases the damaged base leaving a single-strand DNA break intermediate with a 3' end that blocks DNA polymerase synthesis. However, apurinic/apyrimidinic endonuclease 1 (APE1) produces a 3' hydroxyl group (OH) suitable for extension by a DNA polymerase. The trinucleotide repeat (TNR) strand is displaced during gap-filling synthesis and TNRs from the displaced 'flap' can fold back into a hairpin. The hairpin DNA is ligated and expansion occurs after the DNA hairpin loop is incorporated into duplex DNA (McMurray, 2010).

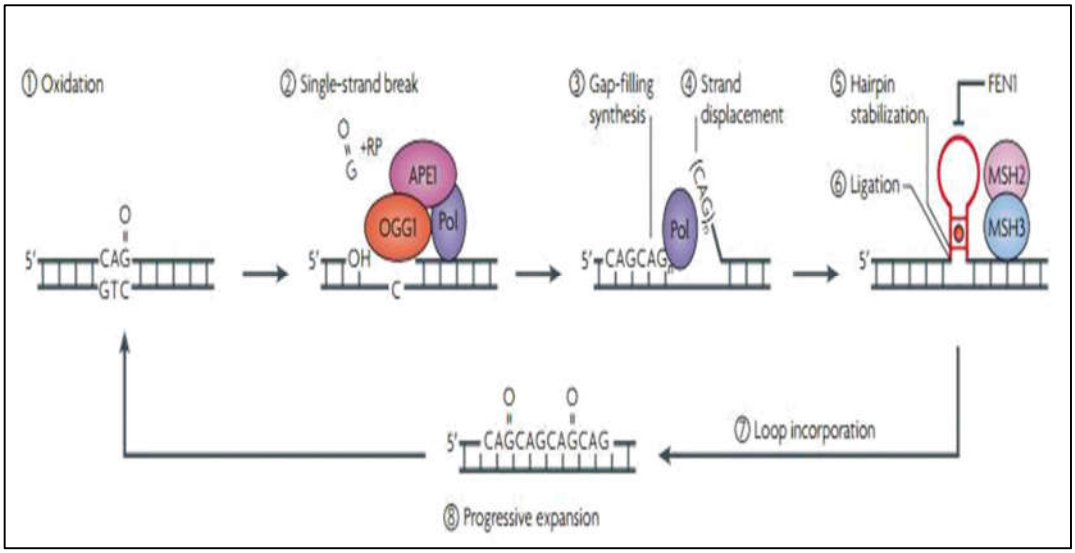


Fig. 4: Loops formed during base excision repair by strand displacement (Source: Mcmurray, 2010)

2.1.4 Open Reading Frame Expansion:

2.1.4.1 Huntington's Disease (HD):

Huntington's disease is a rare neuropsychiatric disorder with a prevalence of 5-10 per 100,000 in the Caucasian population. In Japan, a much lower prevalence of about one-tenth of prevalence in the Caucasian population is described (Bates et al., 2002). HD is a neurodegenerative disorder caused by CAG triplet repeat expansion in *Huntington* (*HTT)* located in chromosome 4p16. This gene is inherited in an autosomal dominant manner with age-dependent penetrance. Alleles in *HTT* are classified as normal, intermediate or pathogenic depending on number of CAG repeats. The normal CAG tract length in the general population is 16 to 29 repeats. Intermediate alleles from 29 to 37 repeats do not cause HD, but are potentially unstable during reproduction. Individuals with more than 37 repeats have an HD causing allele and are considered at risk of developing HD in their lifetime (Pringsheim et al., 2012). Alleles that contain more than 35 CAG repeats are considered HD-causing alleles and confer risk of developing the disease. Alleles that contain more than 40 CAG repeats are completely penetrant. No asymptomatic elderly individuals with alleles of more than 40 CAG repeats have been reported (https://www.ncbi.nlm.nih.gov/books/NBK1305/)

The CAG repeat in *HTT* is translated into an uninterrupted stretch of glutamine residues that when expanded may have altered structural and biochemical properties. The protein product of the *Hungtington* has been shown to form intracellular inclusions and is cytotoxic. Aggregation of the polyglutamine repeat portion of Huntington appears to be associated with the onset of disease and severity of disease (Lutz, 2007).

2.1.4.2 Spinocerebellar Ataxias:

Autosomal dominantly inherited spinocerebellar ataxias (SCAs) are typically present in the middle age and progress over 10 -20 years to cause premature death. Juvenile as

well as late onset cases have been observed owing to larger or smaller CAG repeat expansions, respectively.

2.1.4.2.1 Spinocerebellar Ataxia Type 1 (SCA1):

SCA1 was the first Autosomal Dominant Cerebellar Ataxia (ADCA) Type 1 to be genetically classified in 1974 (Yakura et al., 1974). SCA1 is typically present in the 4th decade of life, although childhood onset and late-adult onset have been reported. In India, SCA1 accounts for 22% of ADCA. This disorder is molecularly characterised by an expanded (CAG)n trinucleotide repeat in the *ataxin-1* located in chromosome 6p22 resulting in an expanded tract of glutamine. *Ataxin-1*, results in variable CAG repeats ranging from 6 – 44 repeats in the general population. Alleles with less than 35 repeats are normal alleles and not been associated with the SCA1 phenotype. They have been found to have CAT trinucleotide interruption(s) that are considered nonmutable. Alleles with the range of 36 – 44 repeats devoid of CAT interruptions are mutable normal alleles. In SCA1 the abnormal protein accumulates in the nucleus as a single aggregate, often referred to as a Nuclear Inclusion (NI). The expanded polyglutamine tract resulting from the CAG expansion results in misfolding of abnormal ataxin-1 resulting in insoluble aggregates. Because these NIs also accumulate, affecting portions of the cell's protein refolding and degradation machinery (chaperones, ubiquitin, and proteasomal subunits), it is thought that impaired protein clearance underlies the pathogenesis of SCA1 and related diseases (Chung et al., 1993).

2.1.4.2.2 Spinocerebellar Ataxia Type 2 (SCA2):

Present in the third or fourth decade (average age 30 years), SCA2 is among three most frequent types of ADCA Type 1 with worldwide prevalence. In 1993, Gispert et al. located the locus of SCA2 in Cuban kindred that mapped to chromosome 12q24.1. It has been reported that SCA2 is exclusively responsible for all ataxia cases in the Indian population. The general prevalence of SCA2 is significantly higher among white SCA pedigrees than in the Japanese (Basu et al., 2000). The gene responsible for SCA2, *ATXN2* which encodes ataxin 2- a cytoplasmic protein which is involved in endocytosis, and modifying ribosomal translation and mitochondrial function (Lastres-Becker et al., 2008). The normal CAG repeat length is less than 31. Alleles with 32 repeats are uncommon and more than 33 CAG repeats are considered as “late onset”. The most common disease causing alleles comprises 37 to 39 repeats. Extreme CAG repeat expansion greater than 200 has been also reported (Whaley, 2011). The widest range of age of onset is observed among individuals with fewer than 40 CAG repeats. Some individuals with alleles of 33 and 34 repeats have had onset after age 60 years. In one study, the presence of 37 repeats was associated with ages of onset ranging from 20 to 60 years. The two normal alleles, which account for more than 95% of alleles in most studies, have 22 and 23 CAG repeats. 32 repeats is an uncommon allele and information is insufficient to classify it as normal or pathogenic (Charles et al., 2007). The CAG expansion codes for a protein that has an abnormally long stretch of glutamine amino

acid residues. The biologic consequence of this abnormal protein is undetermined. However, *ATXN2* transgenic mice had accumulation of ataxin-2 in the cytoplasm with no intranuclear aggregates. In vitro, expression of mutated ataxin-2 causes apoptotic cell death (Huynh et al., 2003).

2.1.4.2.3 Spinocerebellar Ataxia Type 3 (SCA3):

SCA3 which is also known as Machado-joseph disease (MJD) is the most common type of SCA in most populations genetically characterized to date. It is more common in Germany, Brazil, The United States, Portugal and Japan. However, SCA3 is absent in the Italian population. The prevalence of MJD in India (<3%) is much lower than that reported in other Asian populations (Mittal et al., 2005). MJD is an autosomal dominant neurodegenerative disorder caused by a polymorphic CAG repeat expansion in gene *ATXN3* located at chromosome 14q32.1. The code sequence "CAG" is repeated in the *ATXN3*, which produces the disease protein called ataxin-3. Mutated protein is prone to fold abnormally and accumulates in affected brain cells which form abnormal clumps known as inclusion bodies that are located in the nucleus of the cell. While the clumps themselves may not be toxic to brain cells, they do reflect a problem in protein folding that likely affects normal properties of the ataxin-3 protein. A normal variation in the CAG trinucleotide repeat encoding a polyglutamine repeat occurs within exon 10. Trinucleotide repeat length is highly variable in normal individuals with the (CAG)n in different alleles varying from 12 – 44 repeats. In many studies, the distribution of CAG repeat numbers in normal alleles has shown a bimodal or trimodal pattern with peaks around 14, 22-24, and 27 (Rubinsztein et al., 1995). Overall, 93.5% of normal alleles have fewer than 31 CAG repeats. Alleles with 45 to approximately 60 repeats are difficult to categorize because they are rare and may be associated with phenotypes other than that of classic SCA3. Alleles of affected individuals with classic SCA3 phenotype range from ~60 to 87 repeats (Durr et al., 1996).

2.1.4.2.4 Spinocereberal Ataxia Type 7 (SCA7):

SCA7 is one of a autosomal dominant cerebellar ataxias caused due to expansion of trinucleotide CAG repeat in the coding region of *ATXN7* on chromosome 3p14. Normal alleles of *ATXN7* contains 4- 35 CAG repeats and alleles with 28 -33 CAG repeats are meiotically unstable and may are susceptible of having a child with an expanded allele. These intermediate alleles represents a reservoir of chromosomes for expansion and therefore explains the persistence of the disorders in spite of marked anticipation characteristic of SCA7 (Mittal et al., 2005). Approximately 75% of normal alleles have 10 CAG repeats and normal allele with greater than 19 CAG repeats has not been reported till date. The significance of alleles between 19 and 27 CAG repeats are available. 34-36 CAG repeats may be provisionally defined as alleles with reduced penetrance while alleles of greater than 36 CAG repeats are subjected to extreme expansions (Michalik et al., 2004). The SCA7 mutation has been identified in various ethnic groups and geographical regions around the world but its prevalence has been shown to be the

highest among cases of familial ADCAs in South Africa (22%) and 55% among 27 genetically confirmed SCAs in the Scandinavian region. While, the prevalence of SCA7 in Indian population is very low (Faruq et al., 2015).

2.1.4.2.5 Dentatorubral – Pallidoluysian Atrophy (DRPLA):

Dentatorubral-pallidoluysian Atrophy (DRPLA) is a progressive disorder of ataxia, myoclonus, epilepsy, and progressive intellectual deterioration in children and ataxia, choreoathetosis, and dementia or character changes in adults. Onset of DRPLA ranges from before one year age to 72 years. DRPLA occurs with highest frequency in the Japanese population. The age of disease onset ranges from 1 to 60 years (mean age is 28.8 years). The disease is caused by CAG expansion in *ATN1* gene on chromosome 12p13, that results in abnormal protein called atrophin 1, which is widely expressed in neurons. The CAG repeats in normal individuals range from 6 – 35 repeat units and mutable normal alleles exist between 20 – 35 repeat units. Pathogenic alleles have CAG repeats units ranging from 48 to 93 repeat units. Clinical features and the age of onset of this ataxia are significantly correlated with the size of CAG repeats (Whaley et al., 2011). Expanded alleles are fully penetrant except for one individual with a mildly expanded number of CAG repeats (51 repeats) who was asymptomatic at age 81 years (Hattori et al., 1999). CAG repeats larger than 17 repeats are significantly more frequent in the Japanese population than in populations of European origin. Although DRPLA has been reported to occur predominantly in the Japanese, individuals with molecularly confirmed DRPLA have been identified in other populations including European and North and South American (https://www.ncbi.nlm.nih.gov/books/ NBK1491/).

2.1.4.2.6 Spinocerebellar Ataxia 8 (SCA8):

SCA8 is a slowly progressive ataxia with disease onset typically occurring in adulthood. Onset ranges from age 1 to 73 years. SCA8 is associated with expansion of an untranslated CTG repeat in *ATXN8OS* gene and complementary CAG repeat in the *ATXN8* gene. These variations results in expression of a CUG expansion mRNA transcript and a polyglutamine protein, respectively, suggesting toxic gain of function at the protein and RNA levels. Normal individuals have 15 – 50 repeats whereas patients with SCA8 typically have between 70 – 250 repeats. However, it is not yet clear whether repeat sizes ranging from 50 to 70 repeats can be pathogenic or not. In contrast to the ataxias caused by polyglutamine expansions, an untranslated CTG expansion causes SCA8 with molecular similarities to the DM mutation causing myotonic dystrophy, second example of a pathogenic CTG expansion causes a central nervous system disease without the multisystemic features of DM (Todd et al., 2010). The SCA8 form of ataxia is thought to account for 2-5% of autosomal dominant forms of inherited ataxia. Epidemiologic studies of the frequency of the *ATXN8OS*/*ATXN8* expansion have not been performed but prevalence of the expansion and the SCA8 form of ataxia may be especially common in Finland (Juvonen et al., 2005). Although the pathogenic effects of the *ATXN8OS/ATXN8* expansion are not fully understood, several lines of evidence

suggest that SCA8-related CUG expansion transcripts (*ATXN8OS)* cause RNA gain-of-function effects (Ranum et al., 2006).

2.1.5 Non-coding Expansion:

2.1.5.1 Fragile X- associated Tremor and Ataxia Syndrome (FXTAS):

FXTAS is caused by the expansion of the CGG repeat in the 5' UTR of *Fragile X mental retardation1 (FMR-1)* located on the X chromosome. The prevalence of FXTAS is estimated at ~1/4000 in males and ~1/8000 in females. There are also some reports of rare single point mutations & genetic variants as causative factors for the diseases other than expansion of the triplet repeat. Individuals with 5–45 copies of the CGG repeats are unaffected, 55–199 CGG repeats are classified as premutations and > 200 CGG repeats are classified as having full mutations with associated developmental disability. The protein product *FMR1*, Fragile X Mental Retardation Protein is an mRNA binding protein expressed in various tissues and is essential for neuronal and intellectual development. Methylation of CGG repeat occurs once *FMR1* CGG repeat expands to the full mutation. Expanded CGG track is recognised as a CpG island which decreases transcription of *FMR1* resulting in significant removal of FMRP expression (Peprah, 2012).

2.1.5.2 Spinocerebellar Ataxia 12 (SCA12):

Spinocerebellar ataxia type 12 (SCA12) is another autosomal dominant ataxia caused by the expansion of a CAG located within a promoter region of the *PPP2R2B* (chromosomal locus: 5q32) that encodes a regulatory subunit of the brain-specific phosphatase PP2A protein. Therefore, the disease process would originate from an abnormally increased activity of a pro-apoptotic protein. Ages of onset ranges from 8 years to 55 years. In most individuals, symptoms appear in the fourth decade of life. Alleles with 4 – 32 CAG repeats are the normal alleles, ten triplets being the common repeat length. Threshold for complete penetrance is not clear in SCA12, however, in individuals bearing repeat length ranging from 66 – 78 triplets complete penetrance was reported (Cholfin, 2001). More than 35 families from India appears to make SCA12, the second most common form of SCA in India, less prevalent than SCA2 and equal to or slightly more prevalent than SCA1 (Srivastava et al., 2001).

2.1.5.3 Myotonic Dystrophy Type 1 (DM1):

DM1 an autosomal dominant neuromuscular disorder is a multisystem disorder with a highly variable phenotypic expression. It is caused by a CTG trinucleotide repeat expansion in the 3' untranslated region of the *myotonic dystrophy protein kinase (DMPK)* located on chromosome 19q13.3. The number of repeats ranging from 5- 34 are found in unaffected individuals and those that lie in the range of 35- 49 are considered abnormal premutations such that carriers are at risk of having affected offspring with larger repeat length. Based on the number of repeats, age of onset and severity of clinical presentation of DM1 patients are grouped into mild, classical and congenital with repeat size ranging from 50 – 150, 100 – 1000 & > 2000 respectively. Myotonin – protein

kinase (DMPK), a 69-kd serine- threonine protein kinase is closely related to cyclic-AMP-dependent protein kinases and to Rho – binding kinases which may interact with a GTP-binding protein that is regulatory subunit of myosin phosphatase (Dryland et al., 2013).

2.2 Mitochondrial DNA (mtDNA):

Mitochondrial DNA (mtDNA) is located within the mitochondrial matrix of mitochondrion. mtDNA unlike nuclear DNA, is not involved in the majority if not all cellular processes, but only in those which occur inside the mitochondrion such as oxidative phosphorylation and ATP synthesis. The origins of the mitochondria are widely accepted to have derived from a mutual symbiosis between the cells and a bacterium. Human mitochondrial DNA is a double-stranded circular molecule of 16,569 base pairs length (Calloway et al., 2005). It consists of longest non-coding region of molecule between nucleotide 16024 and 575, which is known as “Control region” where hypervariable regions HVS-I and HVS –II are located (Taanman, 1999). The rate of mtDNA diversion is less than 2% per site per million years. High substitution rate in the human mitochondrial genome has been attributed to the lack of proof reading activity in mitochondrial DNA polymerase and because of high concentration of oxidative radicals inside mitochondria (Brown et al., 1982).

Many attempts have been made to characterize the relative mutation rates in mtDNA especially in the two HVS regions. HVS-I and HVS-II have few hot-spots sites that exhibit very high mutation rates. The fast substitution rate makes it possible to distinguish relatively recently diverged population. This is the reason mtDNA is extensively used in population studies (Herrnstadt et al., 2002). The analysis of mtDNA has been a potent tool in understanding the human evolution. However, almost all studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7 % of the mitochondrial genome (Ingman et al., 2000).

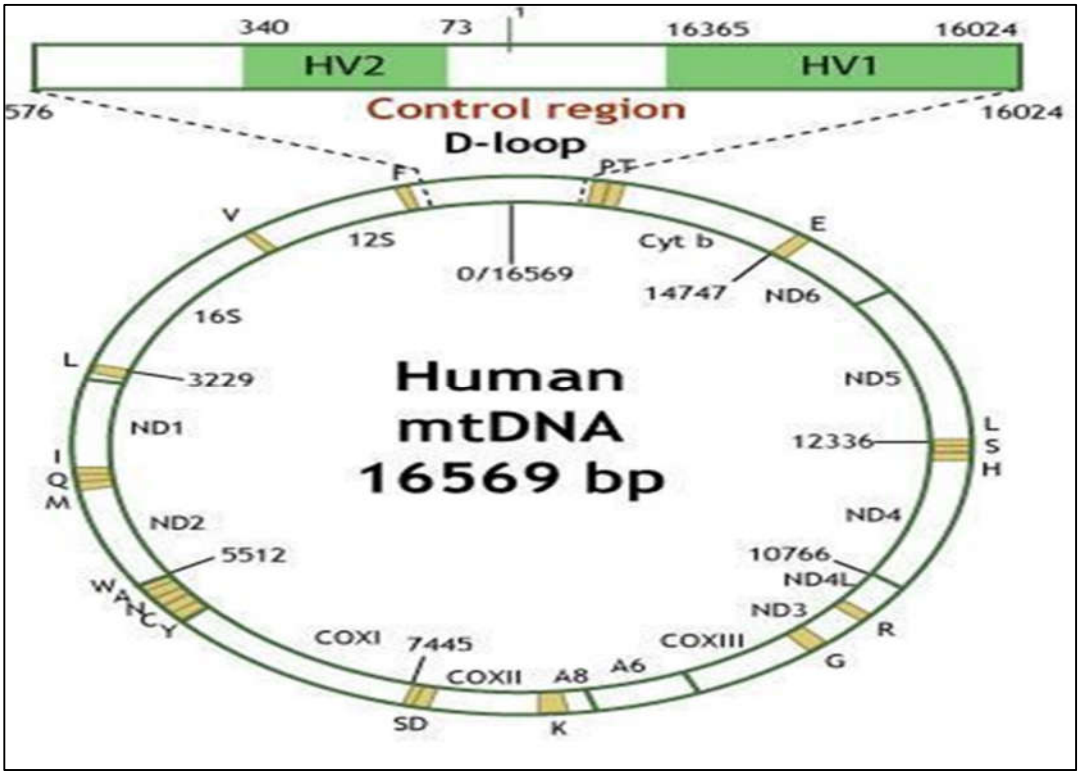


Fig. 5: **Diagrammatic view of mtDNA** showing hypervariable segment I (HV1) and hypervariable segment II (HV2) from position 576 to 16024 along with different component of respiratory chaincomplex (cyt b: cytochrome b, ND6: NADH dehydrogenase 6, ND5: NADH dehydrogenase 5, ND1: NADH dehydrogenase 1, ND3: NADH dehydrogenase 3, ND2: NADH dehydrogenase 2, ND4L: NADH dehydrogenase 4L, COXI, II, III: Cyt *c* oxidase) (Source: http://projects.nfstc.org/)

2.2.1 Mitochondrial DNA Variations:

The nomenclature of mtDNA haplogroups was introduced in the mid-1990s with A-G labels assigned to variation observed in Asian and American lineages, H-K to Europe whereas only a single letter, L, was assigned to describe the highest level of variation observed in Africa in a study using an Asian outgroup (Kivisild, 2015). The likely root of human mtDNA tree is between haplogroup L0 and L1 dividing the phylogenetic tree into two basic clades L0 and the rest as depicted in Fig. 6. L3 was believed to radiate out of Africa, in the form of haplogroup M and N about 60,000 YBP giving rise to the Eurasian variation. Most western Eurasian are characterized by clades within Haplogroup N (Richards et al., 2000).

Quite a number of mtDNA studies focusing on the first hypervariable segment (HVS-I) of the control region (D- loop) have been applied to various Indian populations and have provided some insights into the genetic structure of the populations (Palanichamy et al., 2004). Indian mtDNA lineages belong to either Asian-specific haplogroup M or western Eurasian-specific haplogroups H, I, J, K, U, W. The high frequency and diversity of mtDNA haplogroup M, the major contributor to the Indian maternal gene pool, has been associated with its southwest-Asian origin (Sharma et al., 2005).

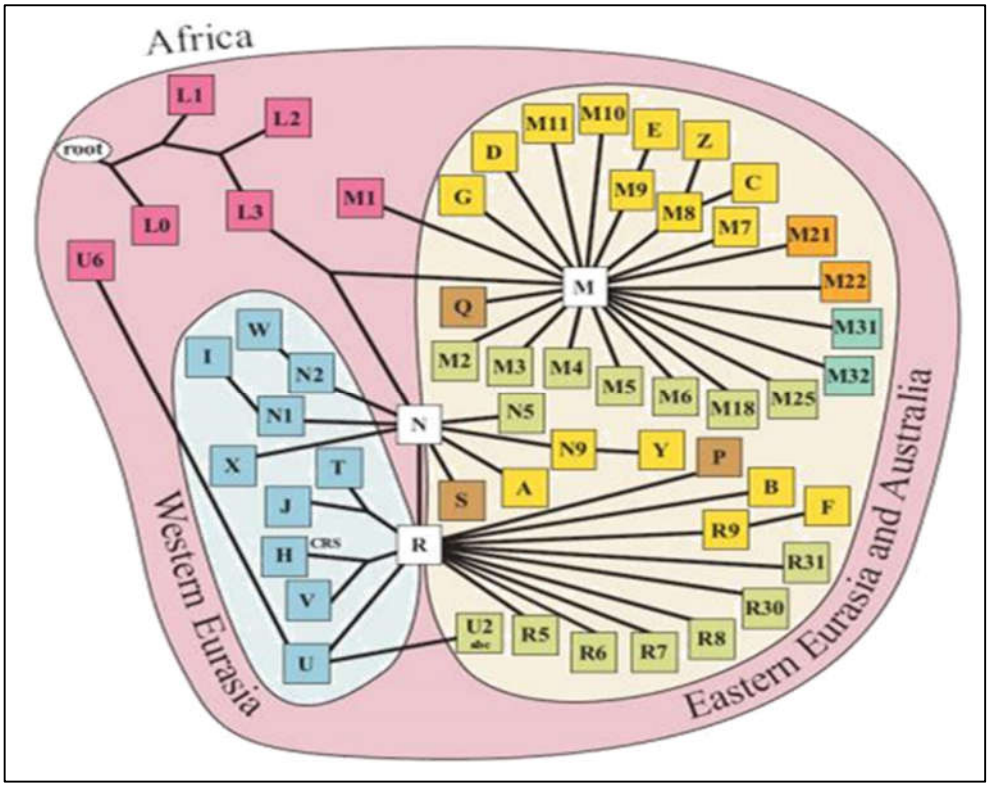


Fig. 6: **A skeleton of the global phylogenetic tree**. Colours: Green – Haplogroups specific for Indian subcontinent, Yellow-Eastern Eurasian Haplogroups, Blue- Westerns Eurasian Haplogroups (Source: Karim, 2005)

The Americas are dominated by five haplogroups A, B and X among native North Americans & haplogroups B, C and D among South Americans. Haplogroups A-D are present in Asia, while X is found in low frequencies outside the Americas. There are a number of minor populations among the larger populations within East Asia. Haplogroups B, F, M7 and R9 tend to be found in abundance within Chinese populations, and in Hong Kong. In Europe, there is one haplogroup in particular which dominates the population landscape, haplogroup H. The frequency of this haplogroup increases the further west into Europe, in addition, there are also regional 'hotspots' where the frequency is greater than the surrounding populations – such hotspots include the Spanish Basques, northern Germany, Denmark, northern France and Great Britain (Whale, 2012).

2.2.2 Out of Africa:

Modern humans originated in Africa more than 200,000 YBP according to the archaeological and fossil evidences. The dispersal of modern humans out of Africa is now widely accepted. Recent findings favour one single dispersal from Africa via southern route, through India and onward to Southeast Asia and Australia. Southern route puts India on the way of earliest migrations of anatomically modern humans. Lack of L3 lineages other than M and N in India suggests that the earliest migrations of modern humans already carried these two mtDNA ancestors. The N branch had given rise to its daughter clade R, which later, in eastern Asians, differentiated into clusters B and R9 (Kivisild et al., 2002).

2.2.3 Maternal Lineages in South Asia:

From the time when mtDNA became one of the important tools in tracing human's prehistoric movements, maternal lineages have been greatly under study. Recently, there has been greater focus on Indian and Asian populations due to their rich anthropological history. In India, the most common lineage is haplogroup M which is also found at high frequencies among the populations inhabiting the region along the southern coast of Pakistan and northwest India. The presence of M haplogroup in Ethiopia, named M1, led to the proposal that haplogroup M originated in eastern Africa. Mitochondrial haplogroup M was arose from African haplogroup L3 about 65000-70000 YBP. But, to contrary Olivier et al. in 2006 reported that about 40,000 to 45,000 years ago North African clades M1 and U6 arose in south-western Asia and moved together to Africa. More than 70 % of present day Indian maternal lineages descends from haplogroup M. Haplogroup U was identified as the common in West Eurasian populations while its 3 subclades U2a, b, and c are present in South Asia which is also the second most frequent haplogroup in Europe (Kivisild, 1999).

2.2.4 Maternal Lineage Study in Nepalese Population:

Most researches taken place to trace origin and migration of different population of Nepal is of Tharu. Tharus are one of the oldest and the largest indigenous people of Terai. Because of its geographic position in a boundary area of Central Asia, Terai was a preferential passageway during the dispersal of many prehistoric and historic populations. Fornarino et al., 2009 in his study revealed a deep common ancestry between Tharus and Indians. In another study by Wang et al., 2012 stated the genetic components of East Eurasian (36.59%) and South Asian (51.63%) ancestry have comprised the vast majority of the Nepalese gene pool. Majority of them belonged to the already defined haplogroups, such as, M3, M5, M18, M30, M35, M43, D4, R8 and M60. The study also revealed that the Nepalese population is closely related to the Tibetan population. The Nepalese lineages of East Eurasian ancestry generally show much closer affinity with the ones from Tibet. There are also studies of Sherpas living around the Himalayas, renowned as high-altitude mountain climber. Bhandari et al., 2015 in their study collected DNA samples of 582 Sherpas living in Nepal and Tibet and studied the genetic diversity in maternal and paternal lineage. In their study they analysed that Tibetans were ancestral population of Sherpas and Sherpa likely acquired their high altitude adaptive features during their ancestor's long stay on the Tibetan plateau before their more recent migration towards Nepal.

CHAPTER 3

MATERIALS & METHODS

3.1 Sample Collection:

Approximately 5 ml of intravenous blood samples from 55 unrelated healthy individuals belonging to Maharjan of Newar ethnic group from Kathmandu valley (Kirtipur, Harisiddhi), Nepal was collected in EDTA-vacutainer tubes. Thirty healthy individuals were from Kirtipur (Kathmandu valley) and twenty five healthy individuals were from Harisiddhi (Lalitpur valley). These healthy individuals comprised of twenty nine males and twenty six females with age group ranged from 21-50 years. All the participants were verbally asked about their health and any related genetic disorder to minimize the chance of participating individuals with any genetic disorder. This research was carried out in Institute of Genomics and Integrative Biology (IGIB), New Delhi, India. A written informed consent was obtained from all the participants and this genetic research has been approved by Nepal Health Research Council, Kathmandu, Nepal.

3.2 Genomic DNA Extraction:

Genomic DNA of 55 unrelated healthy individuals of Maharjan population was extracted by salting out method (Miller et al., 1988). RBC is enucleated, so RBC was first lysed and separated from the whole blood. Nucleus was then disrupted by using Nuclei lysis buffer (NLB) which majorly consists of SDS and NaCl. Once cells were disrupted, nucleic acid along with proteins were released. Amino acids in protein molecules are either hydrophobic or hydrophilic. When salt concentration is increased water molecules are attracted by salt ions. As a result, protein-protein interaction increases and proteins are precipitated out. DNA can be the precipitated using chilled ethanol. The protocol in detail is below.

2-4 ml of EDTA treated peripheral blood was transferred into 14 ml Falcon Tube. About 10 ml of RBC lysis buffer was added into the Falcon and was inverted several times until it became translucent. It was kept at room temperature for about 20 minutes with invert shaking at several intervals. Then, it was centrifuged at 2500 rpm for 10 mins at room temperature (RT). Supernatant was discarded in 20% hypochlorite solution. 15 ml of RBC lysis buffer was added to the pellet and was mixed by brief vortexing. It was centrifuged again at 1000 rpm for 10 min at RT. The supernatant was again discarded in 20% hypochlorite solution. 3 ml of Nucleus lysis buffer (NLB) was added to the pellet and was mixed by vortexing. 160 µl of 10% SDS and 10µl of Proteinase-K (20mg/ml) were added to the tube and were mixed well. Tubes were then incubated at 65^0C temperature for 3 hrs in a water bath. 1ml of saturated NaCl (6M) solution was added and was shaken vigorously for 15 sec. Tubes were spun immediately at 3500 rpm for 35 min at RT. The supernatant was transferred carefully without disturbing pellet and double volume of pre-chilled absolute ethanol (i.e. 100%) was added. DNA was precipitated by inverting

10-20 times very slowly in swirling motion. Precipitated DNA was transferred by using a pipette tip to a micro centrifuge tube containing 1 ml of 70% ethanol. Tubes were then centrifuged at 13000 rpm for 10 min. The supernatant was discarded carefully without disturbing the pellet. Pellet was air dried for 1 hours and re-suspended in 200 µl of TE buffer (Tris + EDTA, pH= 8.0). DNA was dissolved by keeping it at 65^0C for 2 hrs and stored at -20°C.

3.3 Agarose Gel Electrophoresis:

Agarose gel electrophoresis is an efficient technique to separate DNA molecules according to their molecular weights. 0.8 gm of agarose (HI Media) was dissolved in 100 ml 1x TAE buffer and was boiled to dissolve completely. 5 µl of Ethidium Bromide was added from stock solution of 10mg/ml. Gel was cooled down and poured onto a gel tray and was allowed to set. DNA samples were loaded and electrophoresis was carried at a constant voltage of 100V. After 45 min of run halfway the gel was observed under BIO RAD Geldoc.

3.4 DNA Quantification:

DNA quantification was done in Infinite® 200 NanoQuant. 2µl of TE was added in each quartz spot of Nano quant plate and measurement of blank was done. After blank measurement, TE was wiped out by piece of lint free paper and 2 µl DNA samples were applied in the quartz spot of Nanoquant plate. Nanoquant plate were placed in the plate carrier and measurements were performed. Concentration of DNA samples were noted down.

3.5 Polymerase Chain Reaction (PCR):

The polymerase chain reaction (PCR) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence. The technique amplifies specific DNA fragments from minute quantities of source DNA material, even when that source DNA is of relatively poor quality. One DNA molecule is used to produce two copies, then four, then eight and so forth. This continuous doubling is accomplished by specific proteins known as polymerases, enzymes that are able to string together individual DNA building blocks to form long molecular strands. To do their job polymerases require a supply of DNA building blocks, i.e. the nucleotides consisting of the four bases adenine (A), thymine (T), cytosine (C) and guanine (G). They also need a small fragment of DNA, known as the primer, to which they attach the building blocks as well as a longer DNA molecule to serve as a template for constructing the new strand.

3.5.1 Steps in PCR:

1. **Denaturation:** The DNA is denatured at high temperatures (from 90 - 97 degrees Celsius) at which the double stranded DNA gets separated.

2. **Annealing:** Primers attaches to the DNA template to prime extension.

3. **Extension:** Extension occurs at the end of the annealed primers to create a complimentary copy strand of DNA.

3.6 Determination of Trinucleotide Repeats Length Distribution.

Trinucleotide repeats causing tandem repeat disorders were determined by PCR amplification of respective gene. Primer desiging, amplicon size and respective trinucloetide repeat number are mentioned in appendix 5. Amplified DNA were checked in 2% agarose gel & then fragment analysis in ABI 3130xl Genetic analyser was done. Data were analysed by using GeneMapper® software.

Table 3: Primer sequence and amplicon size of 10 different trinucleotide repeat disorders (Source: IGIB)

Disorder	Primer sequence (5'...3')	Tm(^{0}C)	Amplicon size
SCA1	**FP – 5'** CAACATGGGCAGTCTGAG **3'** **RP – 5'** GGTGCGGCCGGTGTTCTG **3'**	55.4 62.1	235bp
SCA2	**FP – 5'** GGGCCCCTCACCATGTCG **3'** **RP – 5'** GTGGCCGAGGACGAGGAGAC **3'**	62.1 64.8	206bp
SCA3	**FP – 5'** CCAGTGACTACTTTGATTCG **3'** **RP – 5'** CTTACCTAGATCACTCCCAA **3'**	54.8 54.8	241bp
SCA7	**FP – 5'** AAGGAGCGGAAAGAATGTCG **3'** **RP – 5'** CACGACTGTCCCAGCATCACTT **3'**	56.8 61.6	298bp
SCA8	**FP-5'** GGTCCTTCATGTTAGAAAACCTGGCT **3'** **RP-5'** CATTCTCAGTCCTCACAAGCCTTTCTCAAA **3'**	65 62.8	309bp
SCA12	**FP – 5'** TGCTGGGAAAGAGTCGTG **3'** **RP – 5'** GCCAGCGCACTCACCCTC **3'**	55.4 62.1	152bp
DRPLA	**FP** – **5'** CTCTTAGCCAACAGCAATGC **3**' **RP** – **5**' GGGGAGGGGTGTGAACAT **3**'	54 60.4	419bp
DM1	**FP** – **5**'AACGGGGCTCGAAGGGTCCTTGTAGC **3**' **RP** – **5**'GATGGGCAAACTGCAGGCCTGGGAAG **3**'	76.1 73.9	171bp
HD	**FP-5'**ATGGCGACCCTGGAAAAGCTGATGAA **3'** **RP-5'**GGCGGCTGAGGAAGCTGAGGA **3'**	72.9 70.5	164bp
FXTAS	**FP- 5'**GCTCAGCTCCGTTTCGGTTTCACTTCCGGT **3'** **RP-5'** CCCGCACTTCCACCACCAGCTCCTCCA **3'**	77.6 79.9	278bp

Table 4: PCR cycling Conditions of 10 different trinucleotide repeat disorders:

Disorders	Denaturation(^{0}C)	Annealing(^{0}C)	Extension(^{0}C)	Cycles
SCA1	95^{0}C for 45sec	58^{0}C for 30 sec	72^{0}C for 30sec	35
SCA2	95^{0}C for 45 sec	62^{0}C for 45sec	72^{0}C for 45 sec	32
SCA3	95^{0}C for 45sec	56^{0}C for 30 sec	72^{0}C for 30 sec	32
SCA7	95^{0}C for 45 sec	65^{0}C for 30sec	72^{0}C for 45 sec	35
SCA8	95^{0}C for 30 sec	60^{0}C for 30sec	72^{0}C for 30 sec	30
SCA12	95^{0}C for 45 sec	57^{0}C for 30 sec	72^{0}C for 45sec	30
DRPLA	95^{0}C for 30 sec	52.6^{0}C for 30sec	72^{0}C for 1min	31
FXTAS	97^{0}C for 35 sec 97^{0}C for 35 sec	64^{0}C for 35 sec 64^{0}C for 35 sec	68^{0}C for 4 min 68^{0}C for 6 min	10 25

DM1	95^0C for 45sec	70^0C for 45sec	72^0C for 3 min	35
HD	95^0C for 45sec	62^0C for 30 sec	72^0C for 45sec	35

3.6.1 Fragment Analysis:

Fragment analysis is a general term used to describe genetic marker analysis experiments which rely on detection of changes in the length of a specific DNA sequence to indicate the presence or absence of a genetic marker. Marker analysis is a general genetic technique in which the sequence of the gene is not directly analysed, but the presence of a particular allele or mutant version of the allele of the gene is inferred from the presence or absence of a linked DNA sequence which can serve as a marker for the allele. Trinucleotide repeats in a given gene can be highly variable, a characteristic which makes them useful as genetic marker. The variation in number of repeats affects the overall length of the gene. The length is determined by PCR using fluorescently labelled forward primers and unlabelled reverse primers that flank both ends of the sequence and thus, generate a DNA fragment whose length depends on the number of the repeats in the sequence. It is these fragments which are analysed.

Analysis is based on principle of capillary electrophoresis. It is a process used to separate ionic fragments by size. During capillary electrophoresis, the extension products of the PCR reaction enter the capillary as a result of electro-kinetic injection. A high voltage charge applied to the sample forces the negatively charged fragments into the capillaries. The extension products are separated by size based on their total charge. Shortly before reaching the positive electrode, the fluorescently labelled DNA fragments, separated by size, move across the path of a laser beam. The laser beam causes the dyes attached to the fragments to fluoresce. The dye signals are separated by a diffraction system, and a CCD camera detects the fluorescence. Because each dye emits light at a particular wave length when excited by the laser, all colours, and therefore the loci, can be detected and distinguished in one capillary injection. The fluorescence signal is converted into digital data, then the data is stored in a file format compatible with an analysis software application (GeneMapper® software).

3.6.1.1 Genescan Protocol:

PCR products were mixed with mixture of Hi-Di Formamide & Genescan marker (Rox-550) in a ratio of 1:9. Denaturation was done in 95°C for 5 minutes and immediately was chilled on ice for about 10 minutes. Samples were then loaded into 384-well plate and briefly centrifuged. Plate was then linked into ABI 3130xl Genetic analyser. Data analysis was done using the GeneMapper® software.

Color	**Base**	**Dye**
Blue	C	5-FAM
Yellow	G	TAMRA
Red	T	ROX
Green	A	HEX

3.7. Identification of Mitochondrial Haplogroup:

Mitochondrial haplogroups were identified by amplification of mitochondrial D-loop region using primer sequence listed in Table 5. D-loop region is approximately 1122 bp that was amplified using 3 primer sets. Primer amplified D-loop sequence is showed in appendix 4. Sequencing of the amplified region was carried out & data analysis was done by DNA Star software.

Table 5: Primer Sequence and Amplicon Size of mt D-LOOP (Source: IGIB)

Primer Name	Primer sequence	Tm(^{0}C)	Amplicon size
M262	FP – CGCTTTCCACACAGACATCA RP – GGGGATGCTTGCATGTGTA	56.8 56.1	447 bp
M15976	FP – TCCACCATTAGCACCCAAAG RP – GGGAACGTGTGGGCTATTTA	56.8 56.8	573 bp
M16413	FP – TGAAATCAATATCCCGCACA RP – GGGTTTGGCAGAGATGTGTT	52.8 56.8	512 bp

Cycling Condition:

	95°C – 30”		**72^{0}C – 10’**
95°C – 5’	**55°C – 30**	**35 cycles**	**4^{0}C - 10’**
	72°C – 30”		

3.7.1 PCR Products Purification:

PCR products were purified by PEG (Polyethylene glycol) to remove the non-specific amplification products, primer dimers or large quantities of unused PCR primers. Polyethylene glycol is a nontoxic water- soluble synthetic polymer. It is used widespread as precipitating agent for the purification of DNA. DNA has highly charged phosphate backbone makes it polar allowing it to readily dissolve in water. When PEG is added to a DNA solution in saturating condition, it forms large random coils in water. This hydrophilic molecule with the right concentration of salt (Na^+) causes DNA to aggregate and precipitate out of solution from lack of solvation. Na^+ shields the negative phosphate backbones causing DNA to stick together. Changing the amount of PEG and salt concentration can aid in size selecting DNA.

3.7.1.1 PEG Purification of PCR Product:

60 µl (double the volume of PCR product) of PEG was added to the PCR plate containing PCR product and was mixed well by vortex. This plate was incubated at room temperature for 10 minutes and then centrifuged at 3200 rpm for 40 minutes. It was inverted gently over the tissue paper & supernatant was discarded by an invert spin till 300 rpm. 100 µl freshly prepared 70% ethanol was added to each tube & was centrifuged at 3500 rpm for 10 minutes. Plate was inverted gently and tapped over the

tissue paper to remove the supernatant. 100 µl 70% ethanol was added again and centrifuged at 3200 rpm for 10 minutes. Plate was again inverted gently over the tissue paper & supernatant was discarded by an invert spin till 300 rpm. Tubes were allowed to air-dry for 20-25 minutes. Pellet was re-suspended in 15 µl of milliQ water.

3.7.2 Sanger Sequencing:

Sanger et al., 1974 used the principles of DNA replication in the development of the process now known as Sanger dideoxy sequencing. This process takes advantage of the ability of DNA polymerase to incorporate 3'- dideoxy nucleotides, nucleotide base analogs that lack the 3'-hydroxyl group essential in phosphodiester bond formation. Sanger dideoxy sequencing requires a DNA template, a sequencing primer, DNA polymerase, nucleotides (dNTPs), dideoxy nucleotides (ddNTPs), and reaction buffer. The Sanger method chain termination reactions are still used, but pouring, running, & reading polyacrylamide gels has been replaced by automated methods. Instead of labelling the products of all 4 sequencing reactions the same (with a radioactive deoxy nucleotide), each dideoxy nucleotide is labelled with a different fluorescent marker. During capillary electrophoresis the fluorescently labelled DNA fragments, separated by size, move across the path of a laser beam. The laser beam causes the dyes on the fragments to fluoresce. An optical detection device on Applied Biosystems genetic analyzers detects the fluorescence. The Data Collection Software converts the fluorescence signal to digital data, then records the data in a *.ab1 file. Because each dye emits light at a different wavelength when excited by the laser, all four colors, and therefore all four bases, can be detected and distinguished in one capillary injection.

Table 6: PCR condition for Sequencing Reaction:

Components	Working Solution (µl)	Steps	Conditions	Temperature	Cycle
Sequencing buffer (10X)	1	1	Initial Denaturation	95^0C for 10 sec	1
Big dye ™	0.8	2	Denaturation	95^0C for 10 sec	40
Forward Primer (10pmol/µl)	0.64		Annealing	55^0C for 10 sec	
Purified DNA	3		Extension	60^0C for 4 min	
MQ	4.56	3	Hold	4^0C for 10 min	
Total reaction volume	10				

3.7.2.1 Sequencing Reaction Purification:

After the completion of sequencing reactions, products have unused primers, dNTPs, ddNTPs, salts etc. In order to remove these unused products purification is necessary before the processing of sequencing plate. Two major reagents were used for the purification.

- **EDTA**: 125mM EDTA is used because of its chelating property. Various ions such as Mg^{2+}act as a cofactor for various enzymes responsible for DNA degradation (eg: DNase). EDTA binds with such ions and prevents the degradation of DNA.

- **Sodium Acetate**: 3M sodium acetate provides Na^{+}. Sodium acetate is used in mixture with absolute ethanol. Dielectric constant of absolute ethanol is less than that of water which allows the interaction of sodium ions with DNA molecules in lesser force thus precipitating the DNA molecules.

3.7.2.2 Purification Procedure

Steps in purification involves chelation, precipitation and washing using EDTA, sodium acetate and ethanol.

1st mix

MQ	10 µl
125mM EDTA	2 µl
Total	**12 µl**

2nd mix

Absolute Ethanol	50 µl
3M NaOAc	2 µl
Total	**52 µl**

12 µl of 1st mix was added to product and briefly vortexed. PCR plate were incubated for 10 min at room temperature. 52 µl of 2nd mix was added to each product & vortexed briefly. Tubes were again incubated at room temperature for 10 min. Tubes were then centrifuged at 3800 rpm for 30 min. Plate was inverted gently over the tissue paper & supernatant was discarded by an invert spin till 300 rpm. 100 µl of 70% Ethanol (for washing) was added and centrifuged at 4000 rpm for 15 min. Supernatant was discarded by inverting the plate on a tissue paper and tapping it slightly. 100 µl of 70% Ethanol (for washing) was added again. Plate was inverted gently over the tissue paper & supernatant was discarded by an invert spin till 300 rpm. Tubes were air-dried for 20-25 minutes. After drying 10 µl of HI-DI was added and kept at room temperature for 5 minutes. Denaturation was done at 95°C for 5 mins and snap chilled on ice for about 10 minutes. Plate was then linked into ABI 3130xl Genetic analyser. Data analysis was done by DNA Star (SeqMan, Edit Man), Chromas. Mitochondrial haplogroup was generated from Haplogrep software.

3.7.2.3 Bioinformatics Analysis:

For sequence analysis of mitochondrial D-loop region DNA star software was used that had two different applications EditSeq and SeqMan. Reference sequence for D-loop was taken from rCRS. With the help of D-loop sequence haplogroup was identified by the help of online software HaploGrep that used D-loop sequences for generating haplogroup. Under bioinformatics analysis Principal Component Analysis was done to get more insight into the affinity of the maternal components observed in Maharjan population. PCA was done using data mining and text analytics software application SPSS var 16.1. Different geographical regions with observed haplogroup frequencies were used from various published research articles and unpublished study (Appendix).

CHAPTER 4

RESULTS

4.1 Subethnic group-Maharjan population:

Healthy Maharjan individuals enrolled for blood samples collection encompassed majorly from the two cities, Kirtipur and Harisiddhi (Lalitpur). Altogether 55 blood samples were collected. During blood collection each individuals were enquired about their health status and any related genetic disorders. 25 and 35 blood samples were collected from Harisiddhi and Kirtipur respectively.

4.2 DNA Extraction and Quantification:

The extracted genomic DNA from blood sample of 55 individuals of Maharjan population was analysed by agarose gel electrophoresis as shown in Fig. 7. The DNA samples were good enough to proceed to PCR amplification. The amplicons were checked in the gel documentation instrument by using BIO RAD geldoc which showed clear bands in all samples (Fig. 8, 11). The DNA was quantified using in Infinite® 200 NanoQuant which showed good amount of DNA

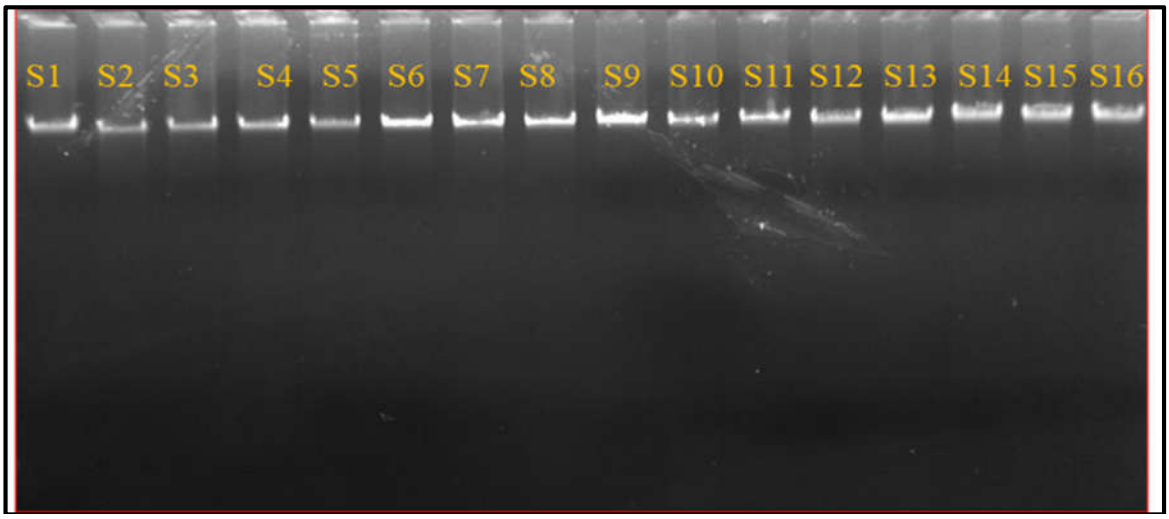
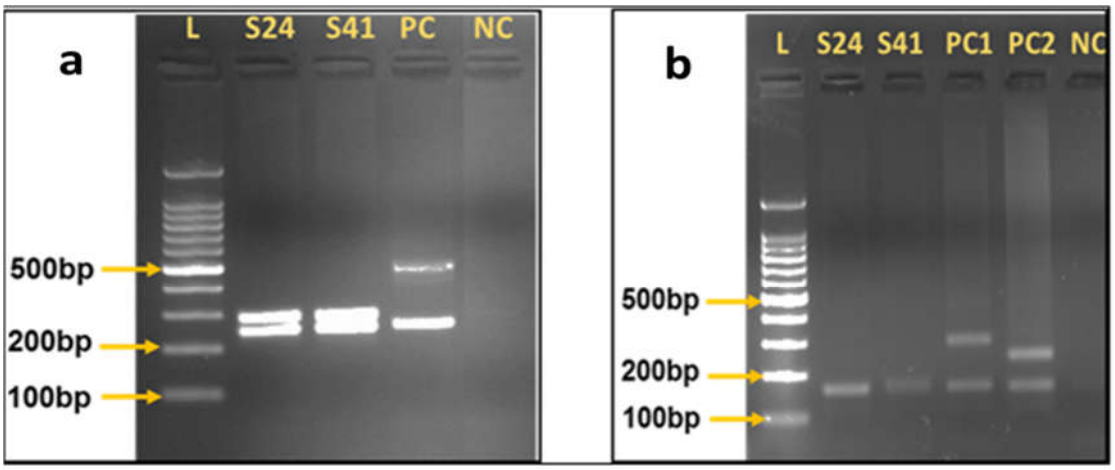
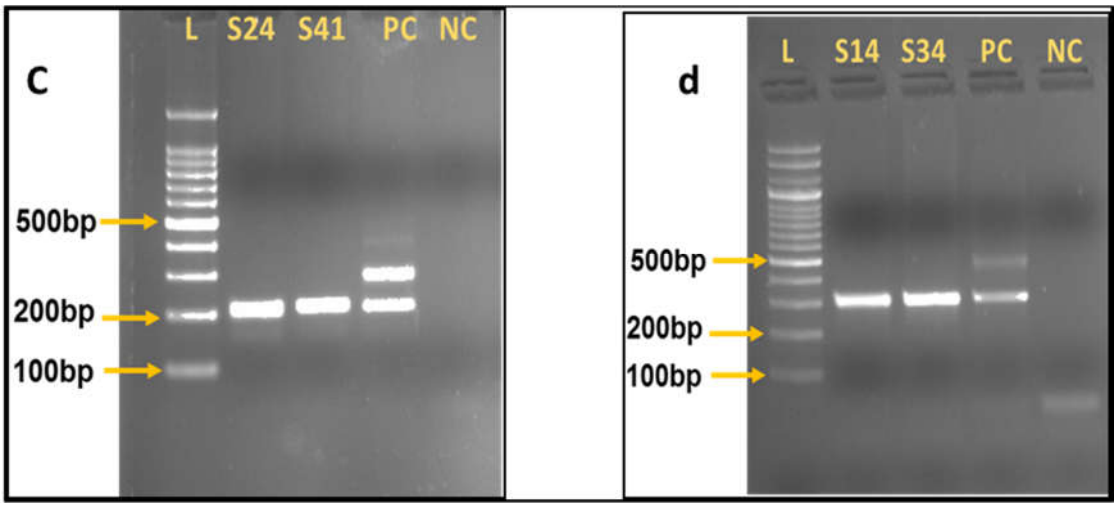


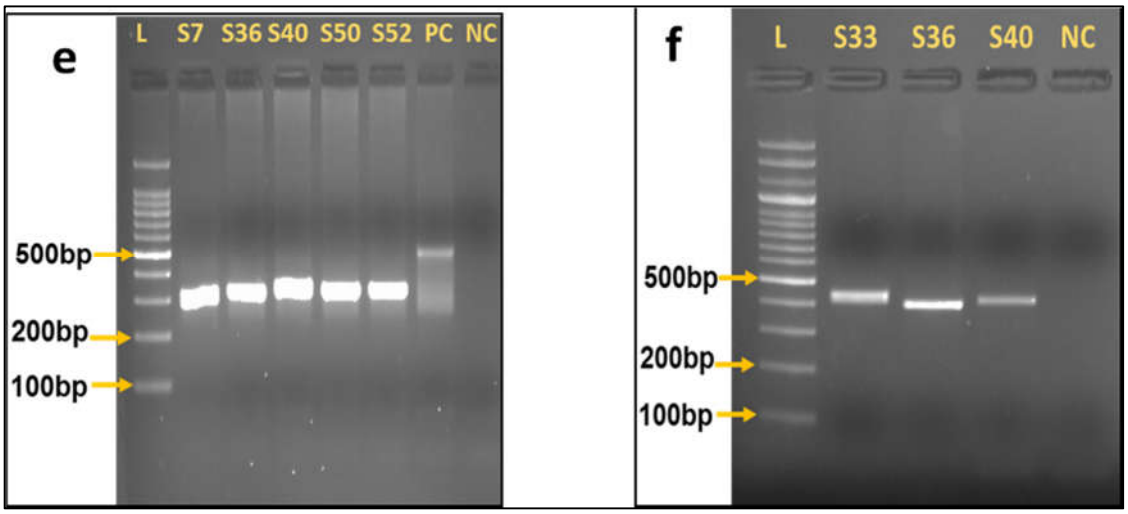
Fig. 7: Agarose Gel Electrophoresis to check diluted genomic DNA for 50 ng in each sample**.** The labels on the top of each well are sample number.

4.3 Optimisation & Detection of Trinucleotide Repeat Length Distribution:

Size of alleles is determined by amplification of specific locus by fluorescent labelled forward primer. Amplified products were electrophoresed on 2% agarose to check the correct amplification of specific locus. All the experimental DNA samples along with the samples of patient as positive control were amplified for comparison between affected and unaffected individuals. Highly expanded alleles of positive control could be visualized with two band due to heterozygosity whereas, experimental DNA samples showed single band or very minimal expanded band of alleles (Fig. 8).


a
L
S24
S41
PC
NC
500bp
200bp
100bp
b
L
S24
S41
PC1
PC2
NC
500bp
200bp
100bp


C
L
S24
S41
PC
NC
500bp
200bp
100bp
d
L
S14
S34
PC
NC
500bp
200bp
100bp


e
L
S7
S36
S40
S50
S52
PC
NC
500bp
200bp
100bp
f
L
S33
S36
S40
NC
500bp
200bp
100bp

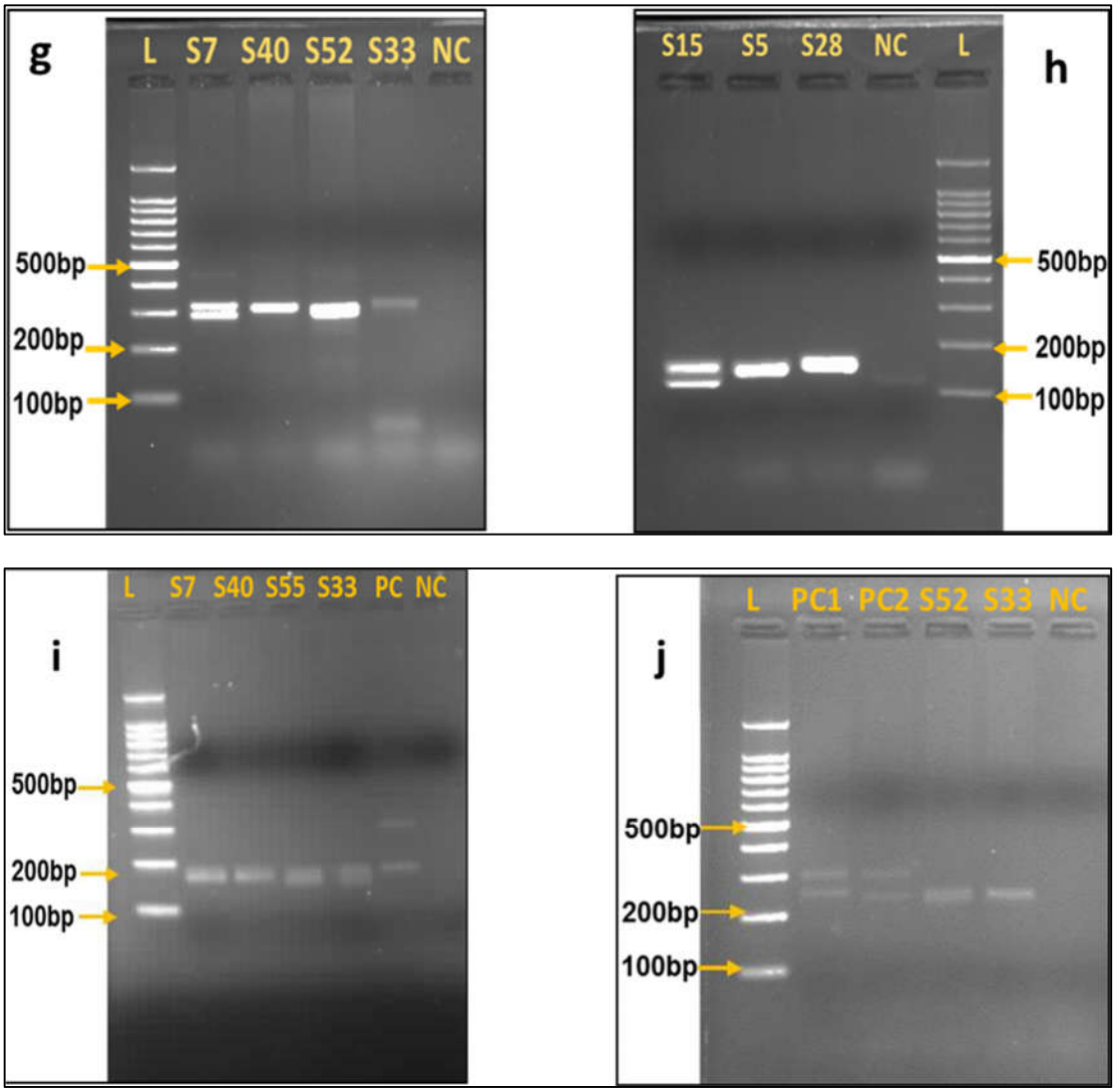
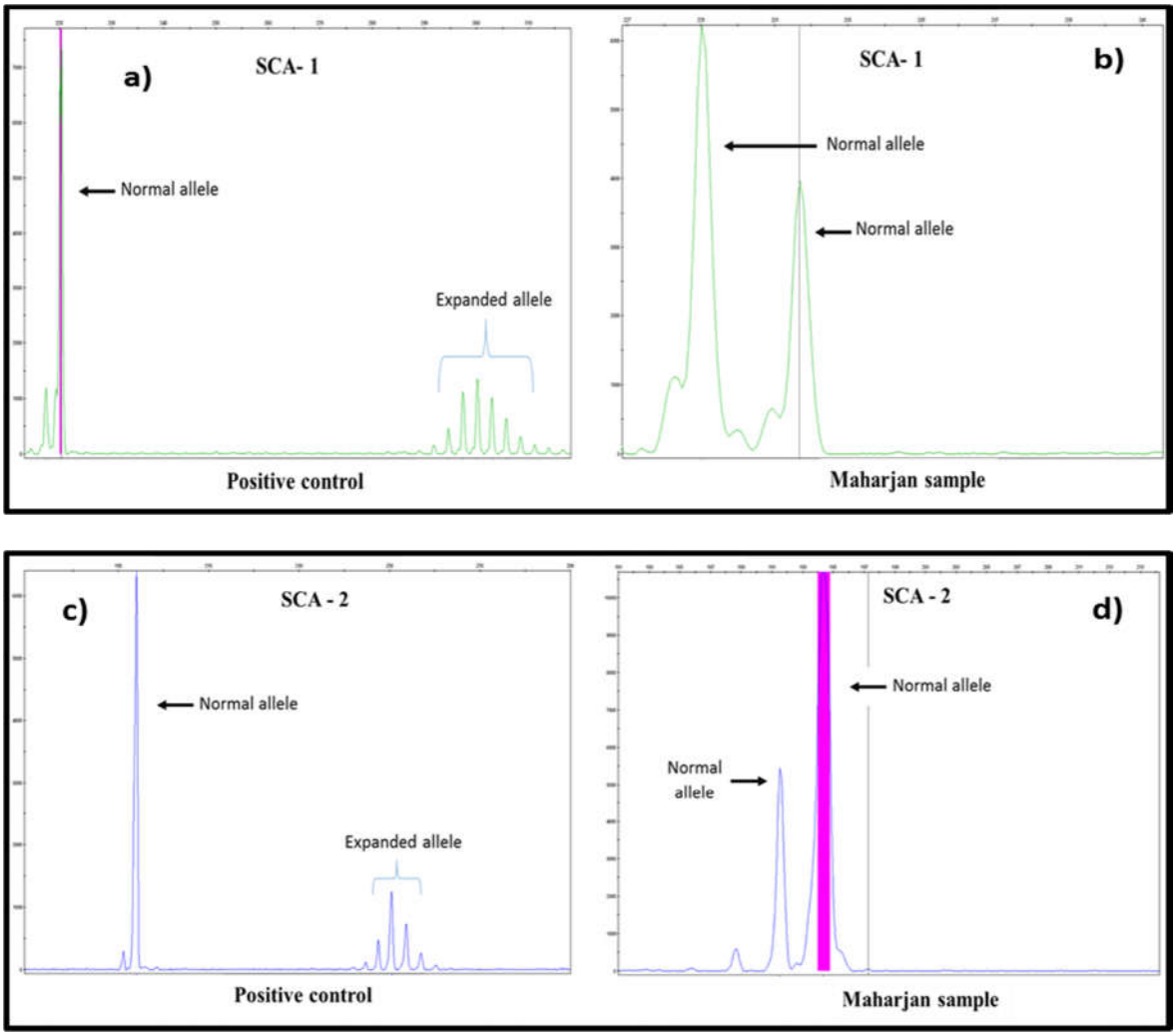
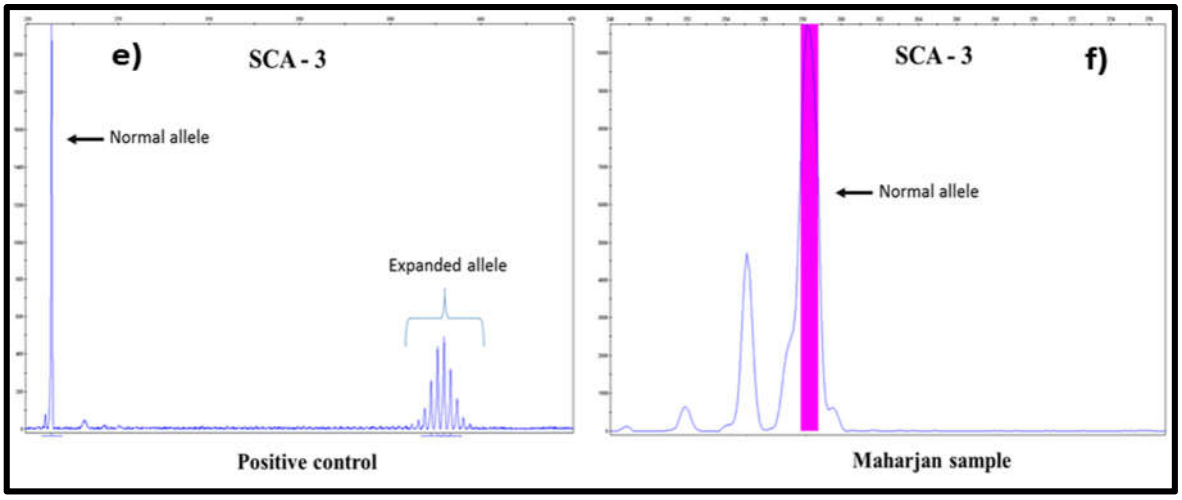


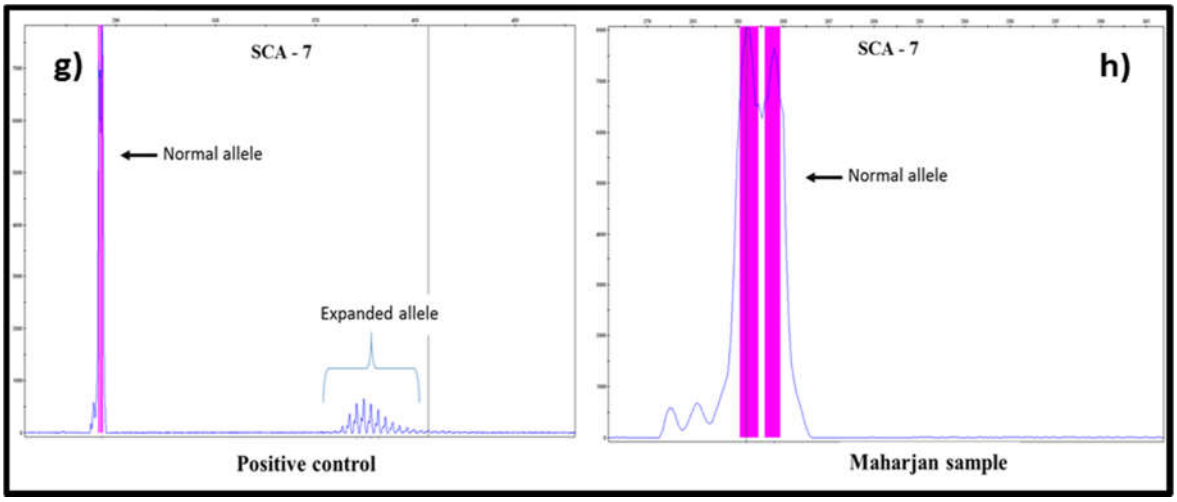
Fig 8: **Gel image of PCR amplicon of various trinucleotide repeat locus in Maharjan samples**. L: 100bp ladder, NC: Negative control; PC, PC1 & PC2: Positive controls **a)** Gel image of PCR product of SCA3; Product size: 241 bp in S24 & S41 **b)** Gel image of PCR product of HD; Product Size: 164 bp in S24 & S41 **c)** Gel image of PCR product of SCA2; Product size: 206 bp in S24 & S41 **d)** Gel image of PCR product of SCA7; Product size: 298 bp in S24 & S41 **e)** Gel image of PCR product of FXTAS; Product size: 278 bp in S7, S36, S40, S50 & S52 **f)** Gel image of PCR product of DRPLA; Product size: 419 bp in S33, S36 & S40 **g)** Gel image of PCR product of SCA8; Product size: 309 bp in S7, S40, S52 & S33 **h)** Gel image of PCR product of DM1; Product size: 171 bp in S15, S5 & S28 **i)** Gel image of PCR product of SCA12; Product size: 152 bp in S7, S55, S40 & S33 **j)** Gel image of PCR product of SCA1; Product size: 235 bp in S52 & S33.

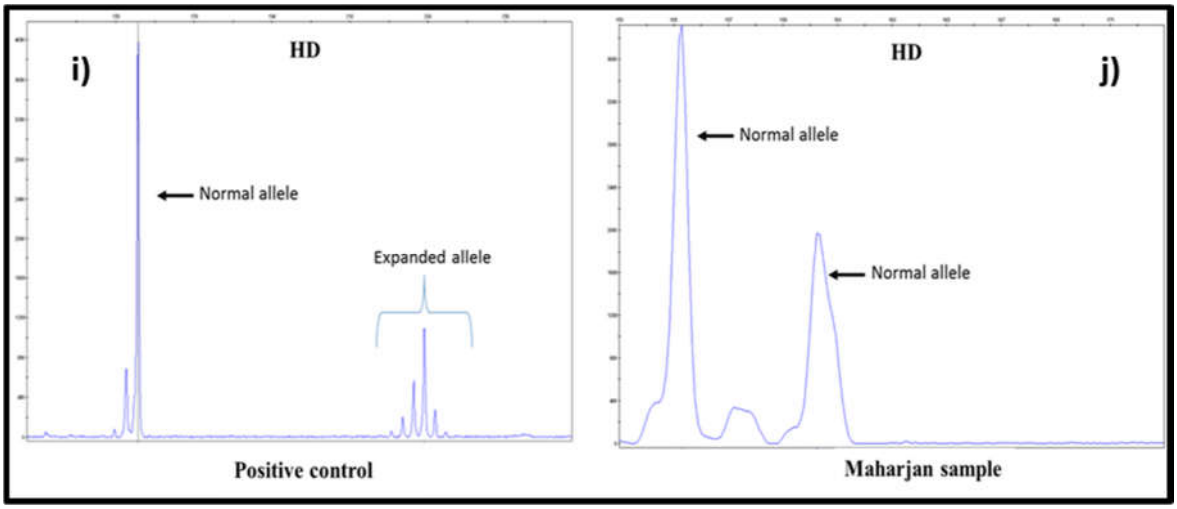
4.3.1 Fragment Analysis to Determine the Amplicon Size by Capillary Electrophoresis:

Expanded alleles can be visualised but accurate size of the amplicon cannot be determined in gel imaging system. To determine the accurate size of the alleles, amplified products were further analysed by capillary electrophoresis. The capillary electrophoresis was carried out in ABI 3130xl Genetic analyser. Fluorescent dyes used for labelling forward primers were FAM and JOE (Thermo fisher SCIENTIFIC). FAM labelled primers showed blue peak (Fig. 9c, 9d, 9e, 9f, 9g, 9h, 9i, 9j, 9k,9l, 9m, 9n, 9p, 9q) whereas primers labelled with JOE showed green peak (Fig. 9a, 9b and 9o). The representative figures have been shown for the details of fragment analysis in which each trinucleotide locus is analysed with a positive control except for SCA8, DRPLA and DM1. Each positive control has one allele with expanded repeats which is visualised by multiple number of peaks while, those experimental samples (denoted as Maharjan sample in the figure) having normal repeats are visualised as single peak. Each peak represents the size of the amplicon.




e)
SCA - 3
Normal allele
Expanded allele
Positive control
SCA - 3
f)
Normal allele
Maharjan sample


g)
SCA - 7
Normal allele
Expanded allele
Positive control
SCA - 7
h)
Normal allele
Maharjan sample


i)
HD
Normal allele
Expanded allele
Positive control
HD
j)
Normal allele
Normal allele
Maharjan sample

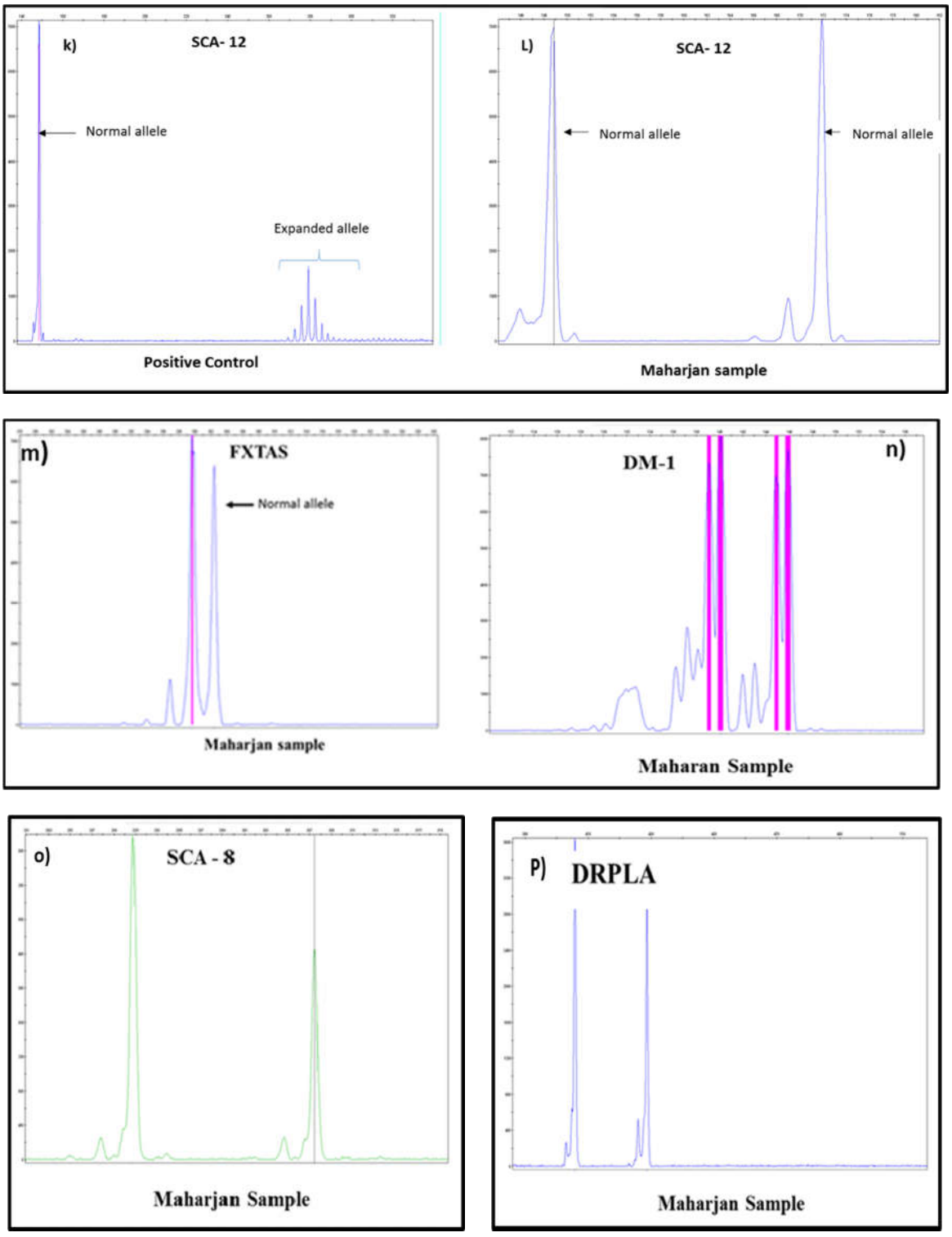
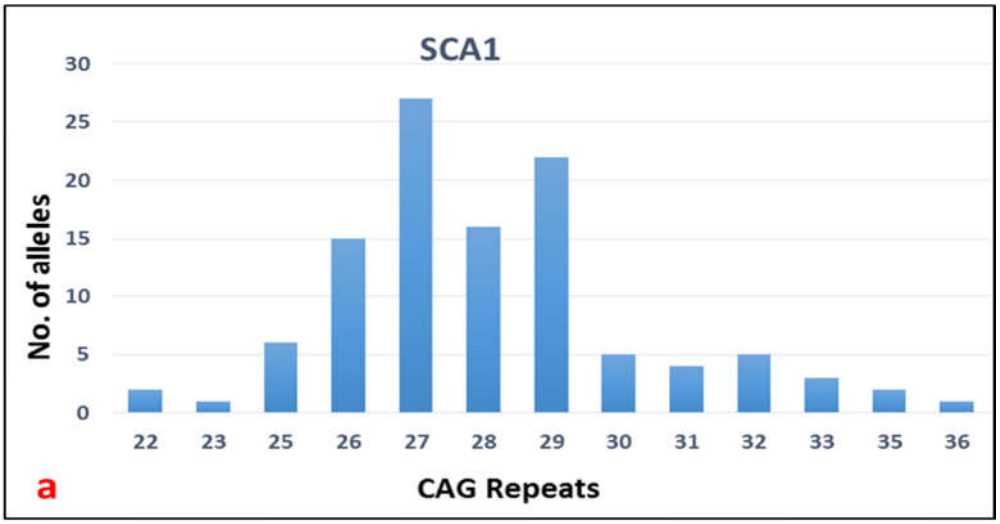


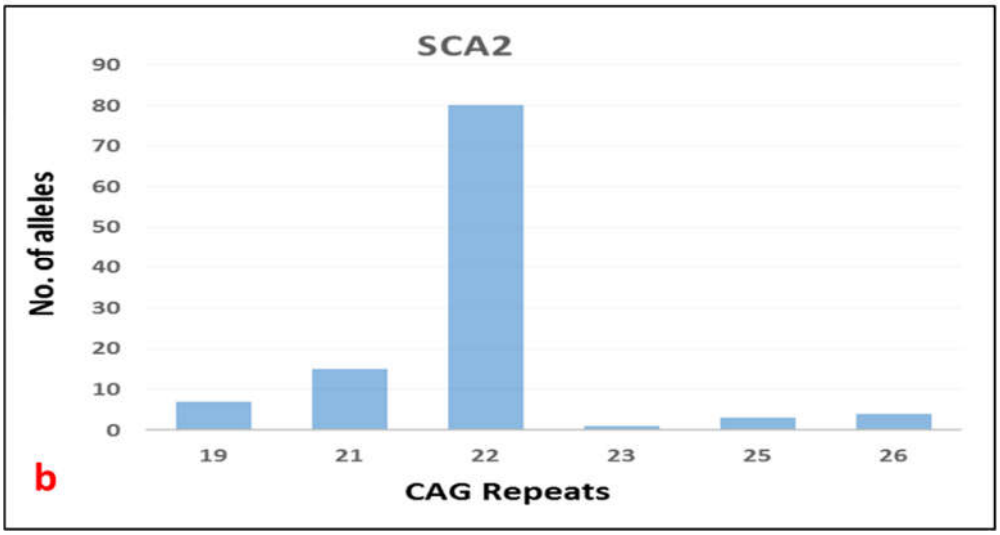
Fig. 9: Chromatogram of fragment analysis. Blue peaks and green peaks observed due to FAM and JOE labelled forward primer respectively. X-axis represents the fragment size & y-axis represents the intensity of the fragment. Two separate peaks represent two alleles of a gene. **a)** Fragment size of normal allele-220.41 bp, expanded allele- 300.10 bp **b)** Fragment size of normal alleles- 148.86 bp and 171.88 bp **c)** Fragment size of normal alleles- 194.04 bp, expanded allele- 250.46 bp **d)** Fragment size of normal alleles- 191.52 bp and 194.34 bp. **e)** Fragment size of normal allele- 240.49 bp and expanded allele- 418.79 bp. **f)** Fragment size of normal alleles-255.10 bp. **g)** Fragment size of normal

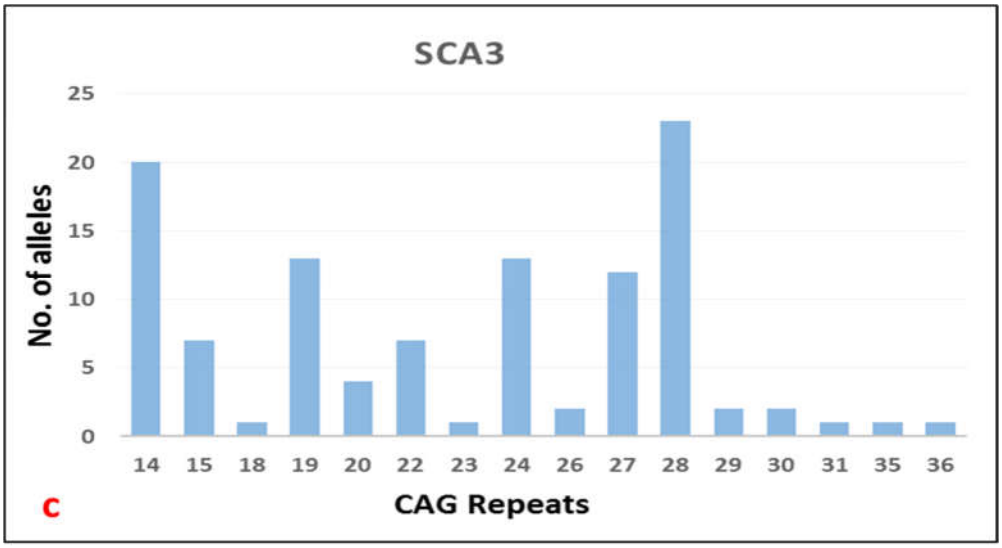
alleles- 284.46 bp and expanded allele- 389.27 bp. **h)** Fragment size of normal alleles- 283.36 bp and 284.59 bp. **i)** Fragment size of normal alleles- 155.68 bp and expanded allele- 229.04 bp. **j)** Fragment size of normal alleles-155.3 bp and 160.24 bp. **k)** Fragment size of normal alleles- 148.53 bp and expanded allele- 279.07 bp. **L)** Fragment size of normal alleles-148.86 bp and 171.88 bp. **m)** Fragment size of normal alleles- 299.56 bp and 302.41 bp. **n)** Fragment size of normal alleles- 140.10 bp and 145.92 bp. **o)** Fragment size of normal alleles- 290.76 bp and 307.47 bp. **p)** Fragment size of normal alleles- 405.80 bp and 428.72 bp.

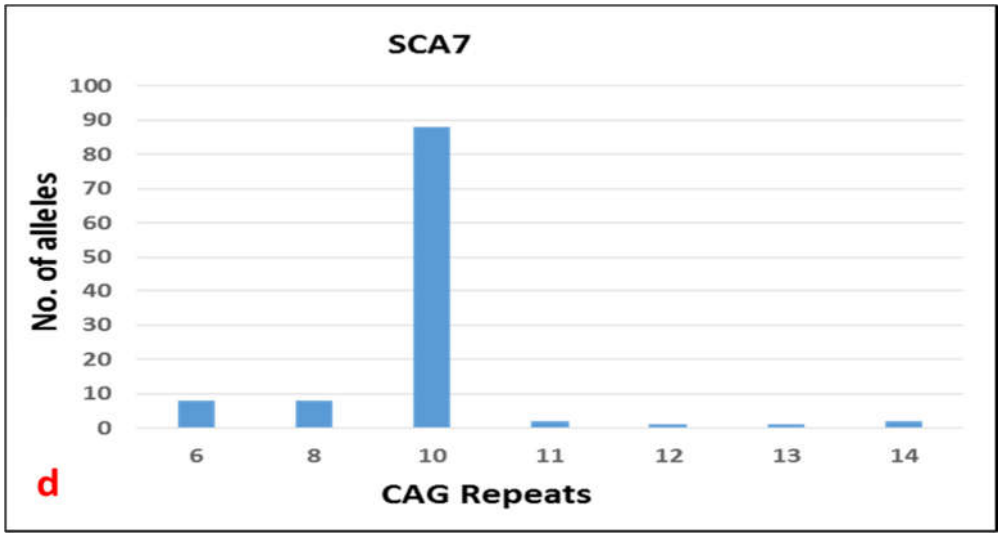
4.3.2: Analysis of Various Trinucleotide Repeats Distribution:

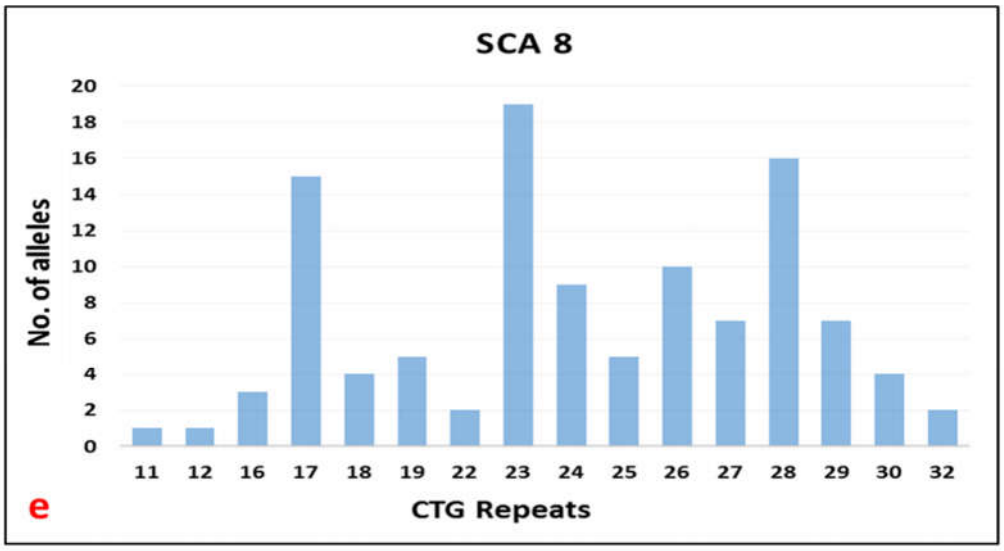
Number of triplet repeats present in wild type sequence were originally determined while primer designing. Each locus of trinucleotide repeat disorders analysed in this study had their different wild type repeat numbers. Amplicon of each ten different trinucleotide repeat locus of each samples were subjected to fragment analysis by capillary electrophoresis. Fragment analysis identified exact amplicon size of each trinucleotide repeat locus in both the alleles. Each observed amplicon size was compared with the expected amplicon size and its repeat and number of triplet repeat in each sample was calculated. SCA1, SCA2, SCA3, SCA12, SCA7, SCA8, DRPLA, DM1, HD, FXTAS loci were analysed in 110 alleles from 55 normal individuals of an ethnic Maharjan population of Nepal. Number of each repeat with their frequencies are represented in bar diagram (Fig. 10).

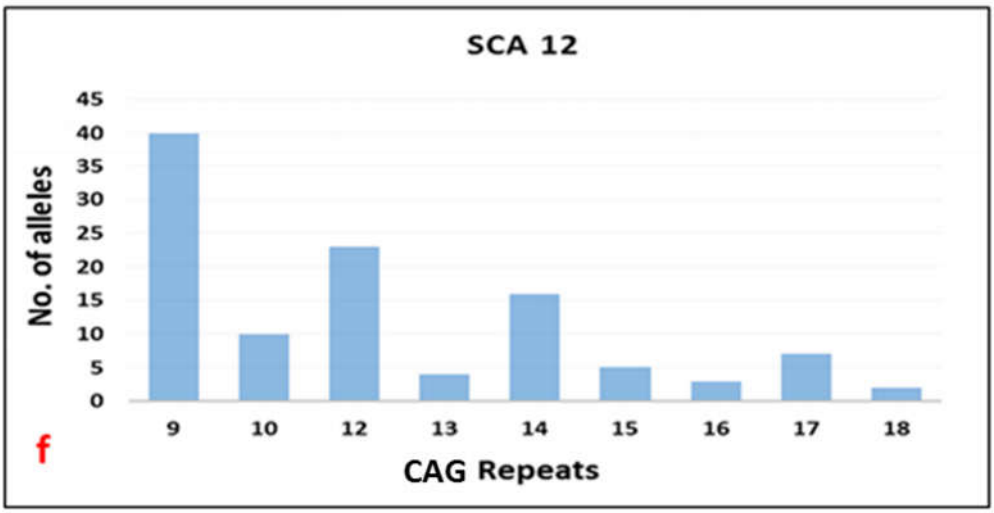


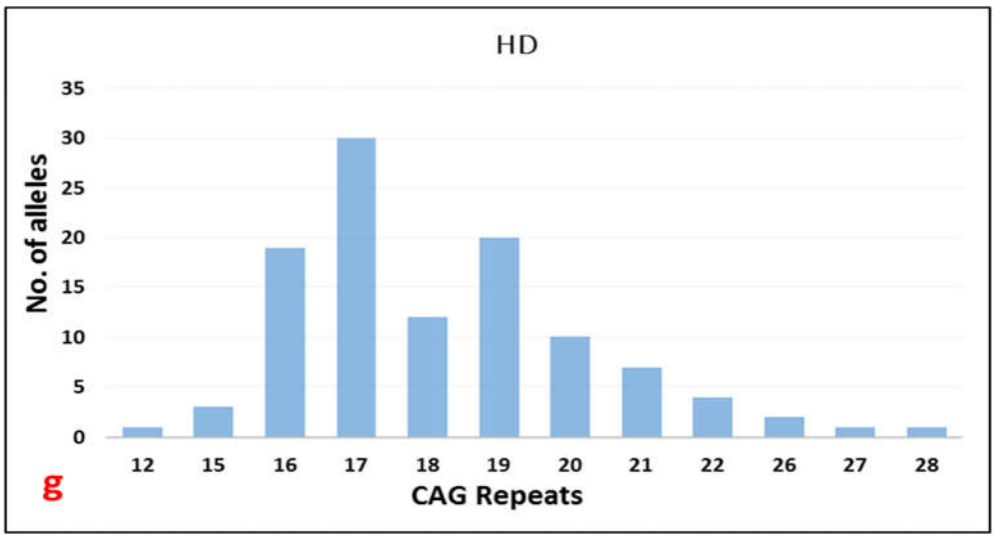

SCA2
No. of alleles
90
80
70
60
50
40
30
20
10
0
19
21
22
23
25
26
CAG Repeats
b


SCA3
No. of alleles
25
20
15
10
5
0
14
15
18
19
20
22
23
24
26
27
28
29
30
31
35
36
CAG Repeats
c


SCA7
No. of alleles
100
90
80
70
60
50
40
30
20
10
0
6
8
10
11
12
13
14
CAG Repeats
d


SCA 8
No. of alleles
20
18
16
14
12
10
8
6
4
2
0
11 12 16 17 18 19 22 23 24 25 26 27 28 29 30 32
CTG Repeats
e


SCA 12
No. of alleles
45
40
35
30
25
20
15
10
5
0
9 10 12 13 14 15 16 17 18
CAG Repeats
f


HD
No. of alleles
35
30
25
20
15
10
5
0
12 15 16 17 18 19 20 21 22 26 27 28
CAG Repeats
g

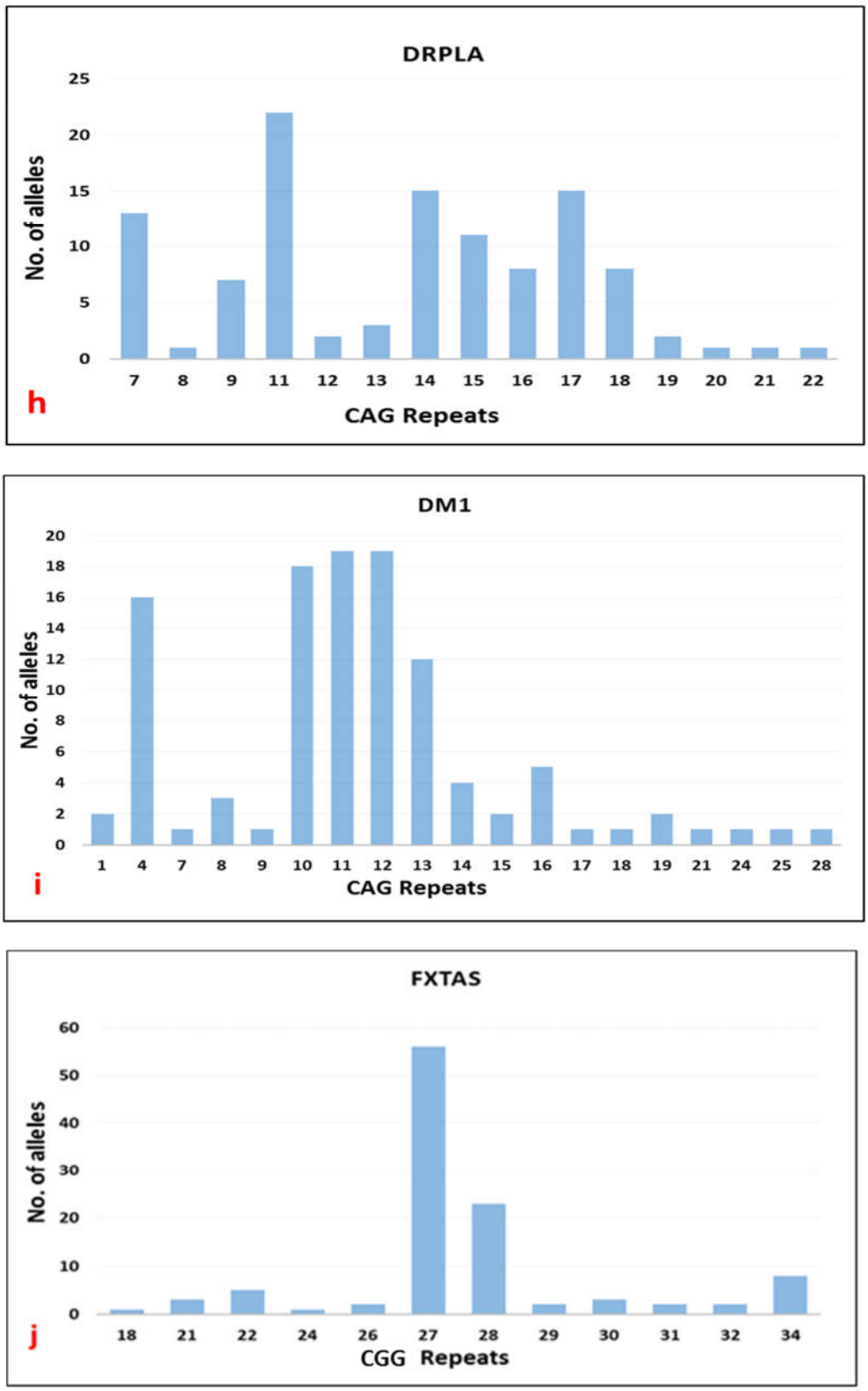


Fig. 10: Graph showing various trinucleotide repeat length distribution among 55 healthy individuals from Maharjan population of Nepal.

Table 7: Trinucleotide repeat range established in this study compared to normal repeat range (McMurray, 2010)

Locus	Triplet Sequence	Location	Repeat range (This study)	Normal repeat range*	Pre-mutation*	Pathogenic*
SCA1	CAG	*ATXN1* (EXON 8)	22-36	6-39	40	41-83
SCA2	CAG	*ATXN2* (EXON 1)	19-26	<31	31-32	32-200
SCA3	CAG	*ATXN3* (EXON 8)	14-36	12-40	41-85	52-86
SCA7	CAG	*ATXN7* (EXON 3)	6-14	4-17	28-33	>36
SCA8	CTG	*ATXN8OS* (3' UTR)	11-32	15-34	34-89	89-250
SCA12	CAG	*PPP2R2B* (5' UTR)	9-18	7-28	28-66	66-78
HD	CAG	*HTT* (EXON 1)	12-28	6-29	29-37	38-180
DRPLA	CAG	*ATN1* (EXON 5)	7-22	6-35	35-48	49-88
DM1	CTG	*DMPK* (3'UTR)	1-28	5-37	37-50	>50
FXTAS	CGG	*FMR1* (5'UTR)	18-34	6-50	55-200	200-4000

* - McMurray, 2010

The allele size for SCA1 locus ranged from 22- 36 CAG repeats, SCA2 locus ranged from 19 - 26 CAG repeats, SCA – 12 locus ranged from 9 - 18 CAG repeats, HD locus ranged from 12 - 28 CAG repeats, DRPLA locus ranged from 7 - 22 CAG repeats, SCA7 ranged from 6 -14 CAG repeats, SCA8 & SCA3 locus ranged from 11-32 & 14-36 CAG repeats respectively. In case of SCA1, 10% of large normal alleles (>31 repeats) from which further expansion of the disease range takes place were observed among the studied population, with higher frequency of 27 repeats. At SCA2 Locus, frequency of large normal allele (>22 repeats) was 7.2%. Large normal alleles (>27 repeats) in SCA3 locus was constituted to be of 27.2% .In case of DRPLA large normal alleles >17 repeats was found to be 11.8%, large Normal alleles of SCA7 (≥ 12 repeats) was 2.7%.

4.4 Analysis of Mitochondrial Haplogroup:

4.4.1 PCR Amplification of mtDNA D-loop Region:

Three different primers were used to amplify the mitochondrial D-loop (M262-M708, M15976-M16548, and M16413-M355). Amplified products were electrophoresed on 2% agarose to check the correct amplification of specific locus as depicted on Fig. 11.

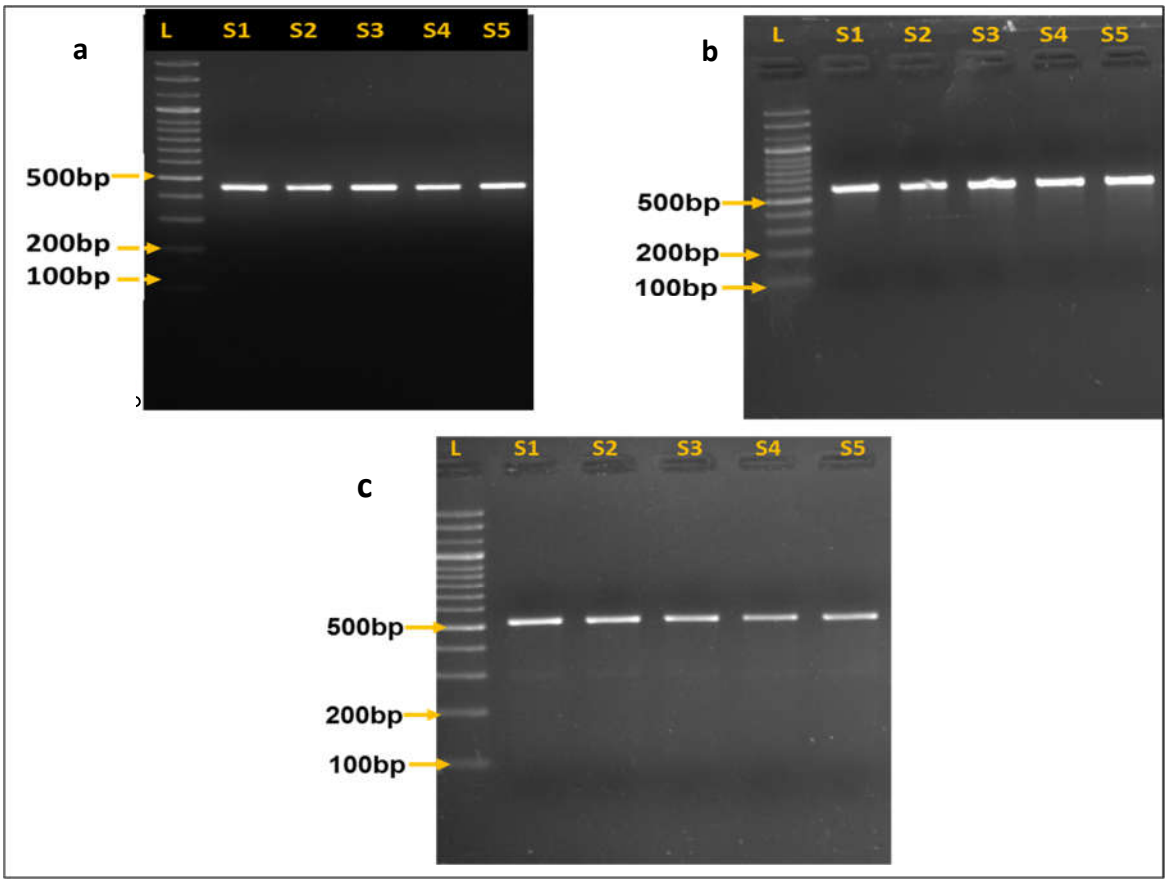
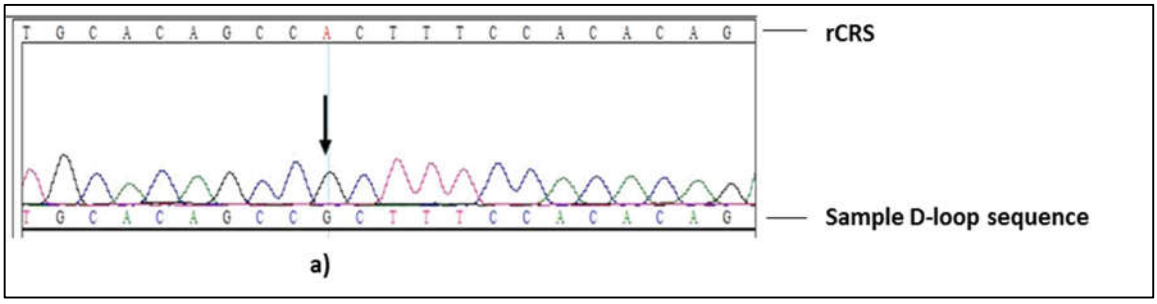


Fig. 11: D-loop amplified region of DNA in 2% Agarose. **a)** Gel image of M262-M708 amplified region of mtDNA; product size- 447, L-100 bp ladder; S1, S2, S3, S4, S5: Maharjan samples. **b)** Gel image of M15976- M16549 amplified region of mtDNA; Product size-573 bp; L-100 bp ladder; S1, S2, S3, S4, S5: Maharjan samples. **c)** Gel image of M16413-M355 amplified region of mtDNA; product size-511 bp; L-100 bp ladder; S1, S2, S3, S4, S5: Maharjan samples.

4.4.2: Sequencing and Aligning with Revised Cambridge Reference Sequence (rCRS)

Sequencing of 1531 bp of mitochondrial DNA that covers the D-loop region of 55 individuals from Maharjan sub-caste of Newar ethnic community, using 3 mitochondrial DNA primers sets, (Appendix) was performed. The sequence of each sample was analysed using a software called DNA STAR. Major mtDNA variations were determined by aligning with rCRS as depicted below in Fig. 12.



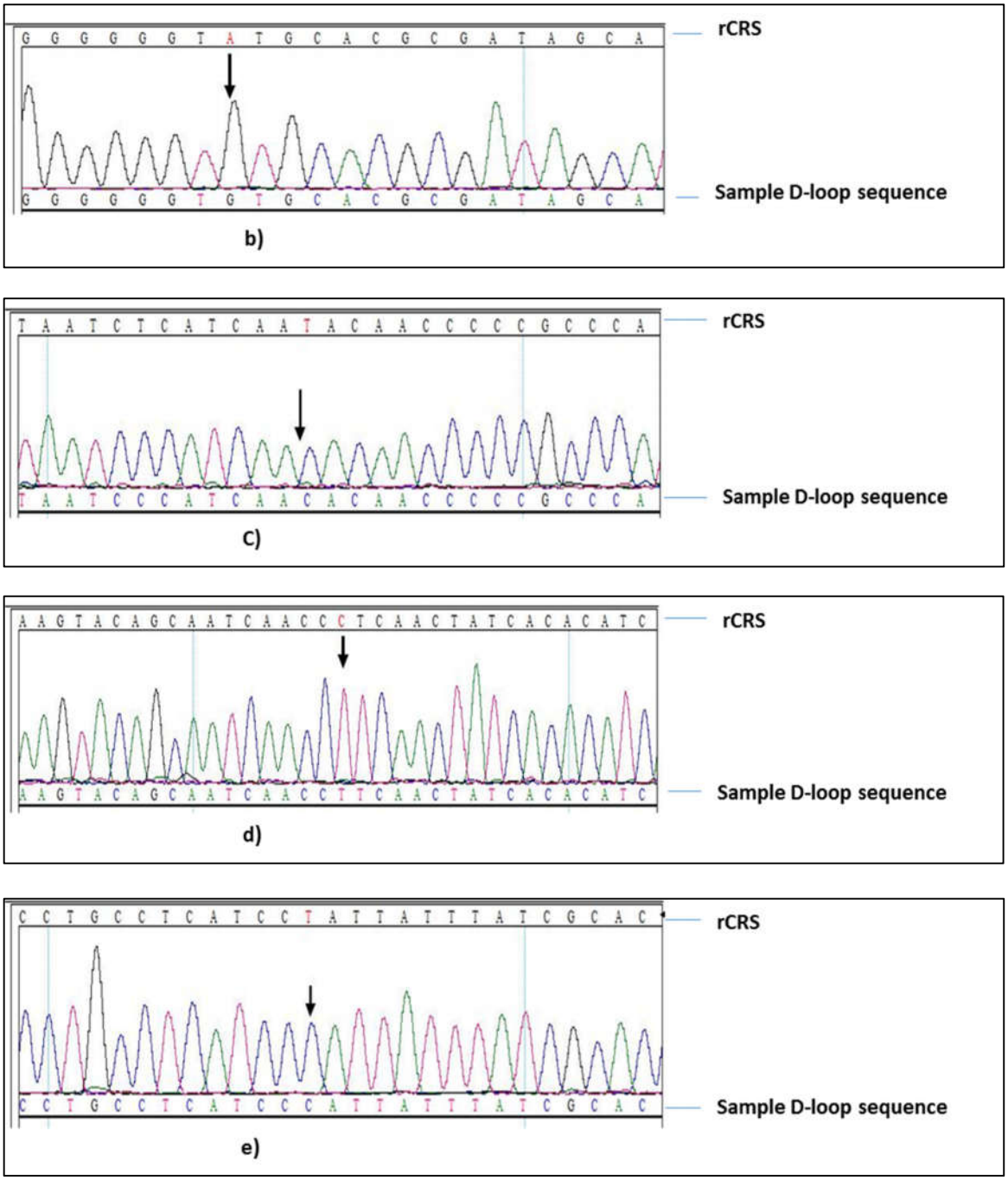


Fig.12: Illustration of sequence analysis of 5 common variations in mt D-loop region of Maharjan population. a) Arrow showing variation at position 73 (A/G). **b)** Arrow showing variation at position 263 (A/G). **c)** Arrow showing variation at position 489 (T/C). **d)** Arrow showing variation at position 16223 (C/T). **e)** Arrow showing variation at position 152 (T/C).

Table 8: Summary of most common variation observed in mitochondrial D-loop region of 55 control individuals of Maharjan population. The most common variation was at position 73 (A/G).

S.N.	Mitochondrial D-loop Variations	Frequency (%)
1	73G	100
2	489C	60
3	146C	30.9
4	152C	33.7
5	16223T	72.7
6	263G	84

4.4.3: Mitochondrial Haplogroup:

Mitochondrial haplogroups were studied in all 55 individuals of Maharjan population. mtDNA variations were identified using a software Haplogrep and a total of 101 different variation were observed in D-loop region. Haplogroups were assigned to each sample based on the polymorphisms on the control region of mtDNA (D-Loop) (Table 8). Major Haplogroup identified were haplogroup M, N H, R & U.

Table 9: Major Haplogroup observed with their frequency among 55 individuals of Maharjan population

S.N	Major Haplogroups	Geographical region	Subclades	Haplogroup Frequency
1	M	South Asia, Central Asia, East Asia	M5, M33, M12'G, M4'67, M3, M24'41, M10, M62'68, M9, M34'57, M7, M8, M31.	58.2%
2	N	Western Eurasian, Central Asia	A,S	18.2%
3	H	East Asia, Western Eurasian, Central Asia	H3,H32	3.6%
4	R	Western Eurasian, East Asia	R9,B4,R2'JT	12.7%
5	U	Western Eurasian, South Asia	U7,U8	7.3%

4.4.4: Mitochondrial Haplogroup Frequency:

Total of 25 different mitochondrial haplogroups were found among 55 healthy Maharjan population. The mitochondrial haplogroup M was found to be most frequent contributing about 58.2% whereas haplogroups N, H, R and U were found to be contribute 18.2%, 3.6%, 12.7% and 7.3% respectively. Haplogroup M is further sub

divided into various subclades (M5, M33, G, M30, M37, M3, M24, M10, M62'68, M9, M34, M7, D, Z, M31). Sub-haplogroup M3 and M33 is observed in higher frequency with 7.2% and 10.9% respectively which is considered to be originated in South Asia from 60,000 YBP. Similarly, haplogroup M is further subdivided into haplogroup D as its descendant which was observed with frequency 7.2%. Haplogroup D was originated in East Asia before 40,000-60,000 YBP. Haplogroup N was observed in highest frequency similar to that of haplogroup M33. Haplogroup N is considered to be predominant in Western Eurasia however, haplogroup M is absent in this geographical region. Haplogroup F and U which is descendants of haplogroup R9 and R respectively was observed in our study population with frequency 9.09% and 7.2%. Other haplogroups M5, M24, M10, M62'68, M9, Z, M31, A, B4, H32, H3, H2a2a2 were observed with frequency 1.8%, G, M30, M37, M with 3.6% and S with 5.4%.

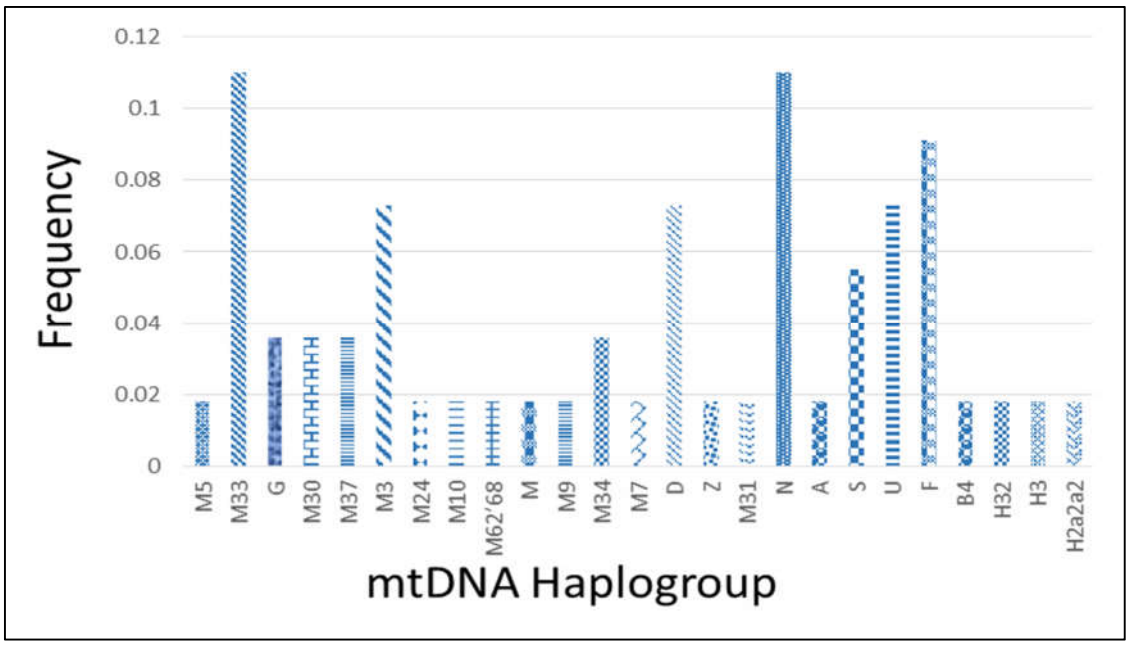


Fig. 13: Graph of Mitochondrial Haplogroup Frequency

4.4.5 Principal Component Analysis of Maharjan population based on mtDNA Haplogroup Frequency

More insights into the affinity of the maternal components were observed in Maharjan population by principal component analysis (PCA) plot. The PCA showed the comparative relationship according to mtDNA haplogroup frequencies with reference to previously studied 26 neighbouring population by using software SPSS var 16.1. Principal Component Analysis transformed the large no. of linear data of different variable into two dimensional simplified data metrics called components which gave better idea about correlation of different variable. Maharjan population was seen comparatively clustered with the Altaian Kazakhs, Uzbeks, Tharu of Eastern Terai and Kathmandu population.

Maharjan population seems lies in the middle with almost equivalent distance between East Asian population (Altai, Uzbeck, Mangolian etc.), Indian populations (South Indian and North Indian and tribe Andrapradesh), and population of Nepal. PCA also shows Maharjan population lies in closer distance with South Asian population like Kathmandu, Newar (Gaydan et al., 2007, 2013) and Nepal_mix, Nepal_Kat (Wang et al., 2012). Our population under study also shows some affinity towards Udaya, Manandhar, Shakya and Bajracharya (Shrestha et al., 2013). This diverse affinity is because Maharjan population under study shares its haplogroups with all other population which were taken for comparison. It suggest that Maharjan population may be the formed by accumulation of genetically, geographically and ethnically diverse population during the course of the history.

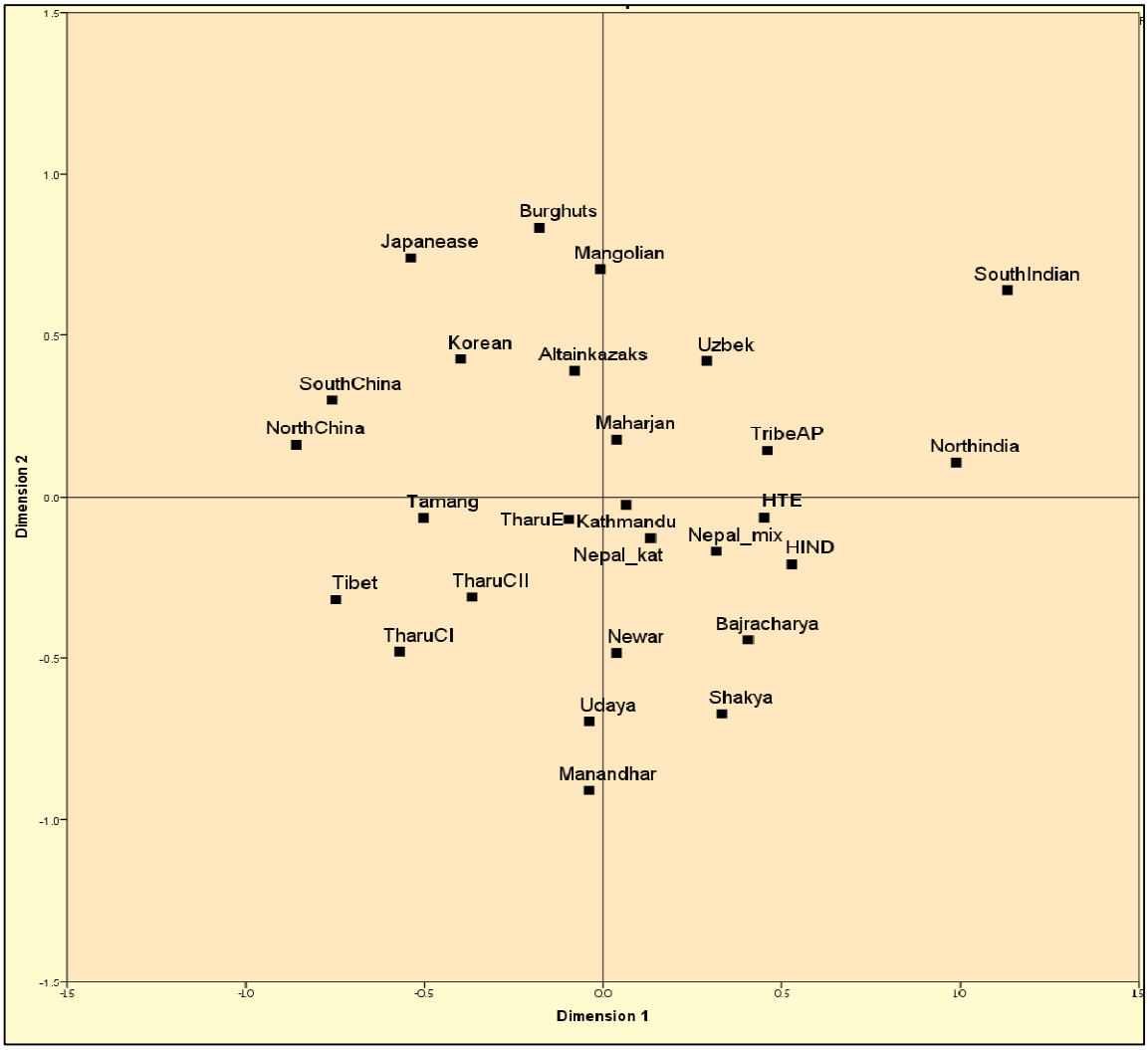


Fig. 14: Principal Component Analysis (PCA) plot of Mitochondrial Haplogroup (where HIND = Hindu India, HTE = Hindu Terai, TribeAP = Tribe of Andra Pradesh, TharuE = Tharu Eastern, TharuCI = Tharu Chitwan I, TharuCII = Tharu Chitwan II, Nepal_kat = Nepal Kathmandu and Nepal_mix = Nepal Mix)

4.4.6: Mitochondrial Phylogenetic Tree of Maharjan Population

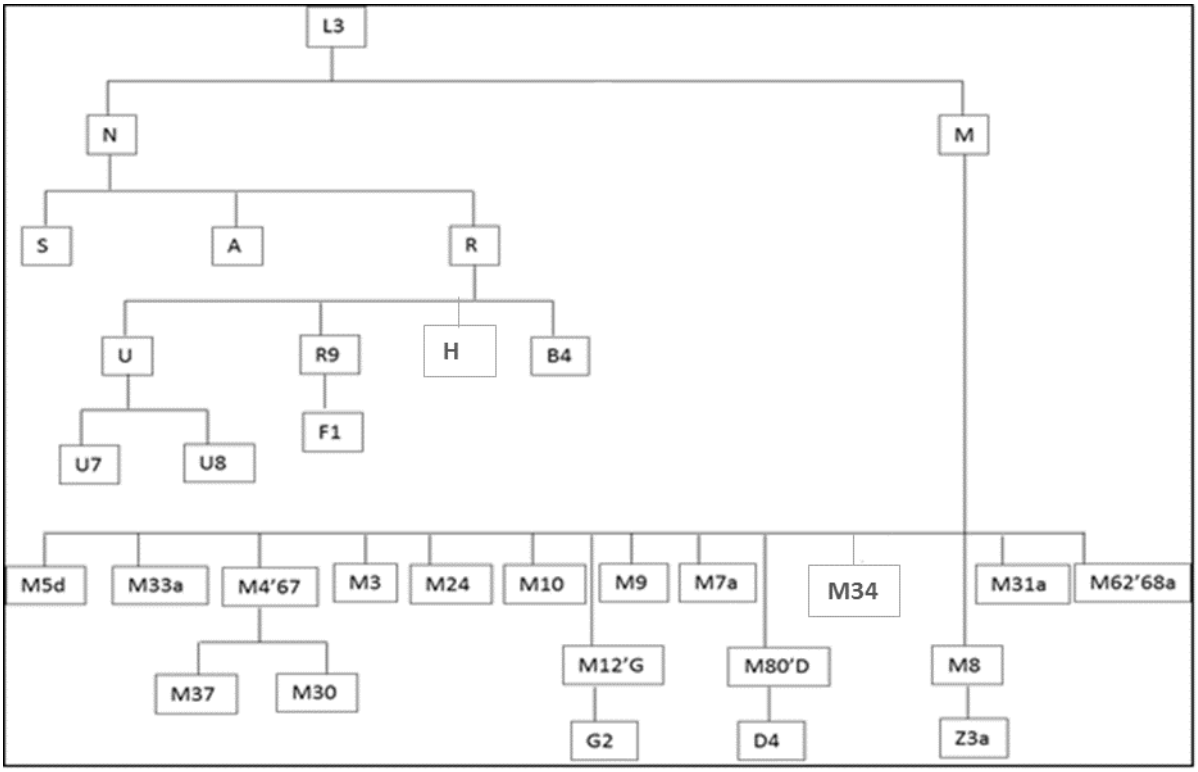


Fig.15: Mitochondrial Phylogenetic tree of Maharjan population according to mitochondrial Haplogroups

L3 is the African haplogroup which was first migrated out of Africa. The macrohaplogroup M which is restricted to Asia, most frequently in South Asia, and the Eurasian macrohaplogroup N were directly descendent from L3. The N phylogenetic node split into A, R and S haplogroups (which are western as well as eastern Eurasian haplogroups) and continued to divide further into specific haplogroups. In our studied population, most of haplogroup were associated with Macrohaplogroup M.

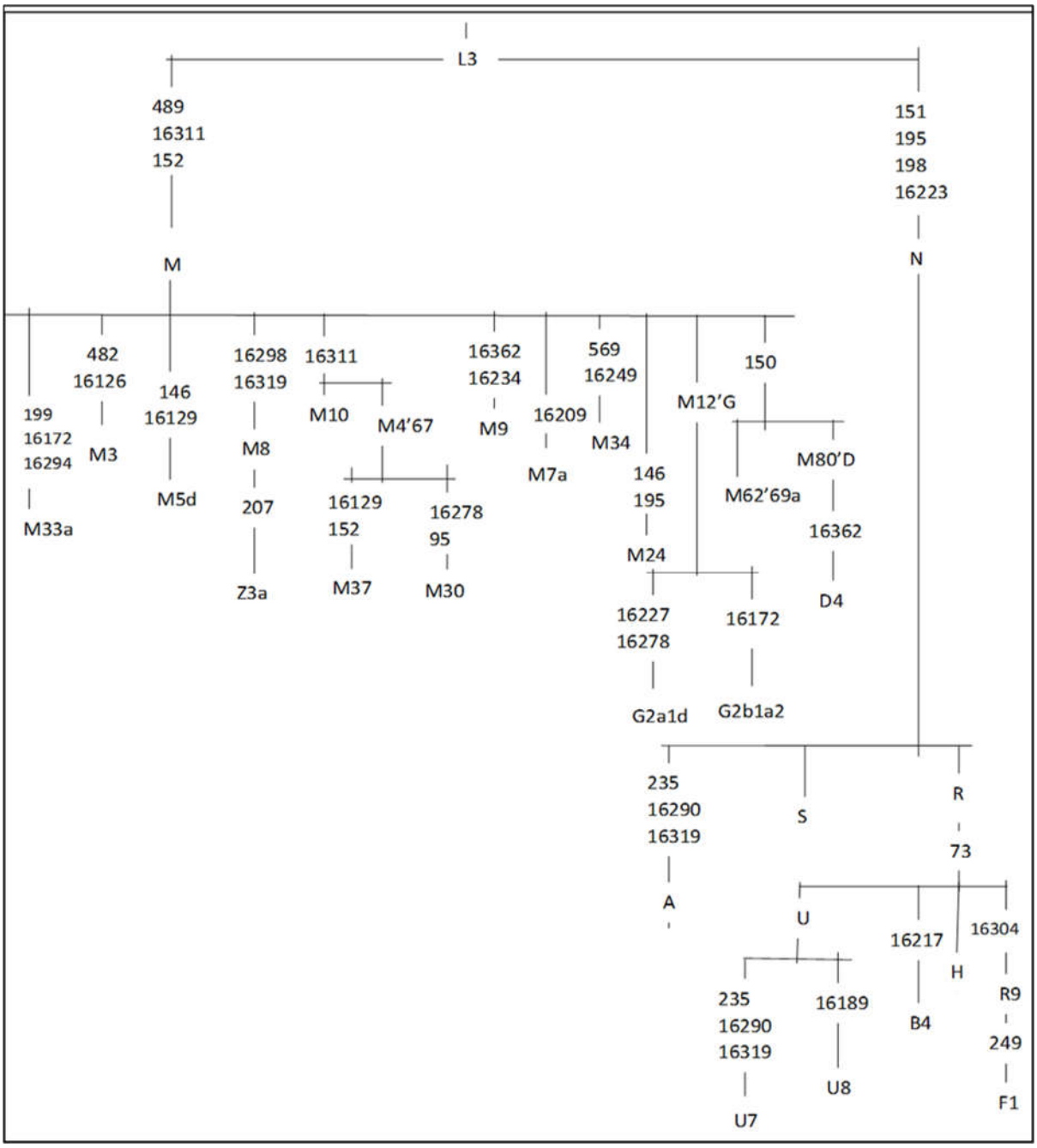


Fig. 16: Phylogenetic tree based on mtDNA D-Loop sequence for haplogroups in Maharjan population. The number along links refer to substitutions scored relative to the rCRS (Andrews et al., 1999)

Table 10: Trinucleotide repeat distribution and mtDNA haplogroups in different individuals of Maharjan population.

Sample	SCA 1	SCA 2	SCA 3	SCA 7	SCA 8	SCA 12	FXTAS	DRPLA	DM1	HD	Haplogroups
S1	26	25	22	10	26	14	27	19	16	17	M62'68
S2	29	22	20	10	24	12	27	17	16	22	M30
S3	28	22	28	10	29	14	28	16	13	20	N
S4	31	22	29	10	12	12	28	14	10	21	N
S5	28	22	27	10	29	14	26	17	12	20	H
S6	28	22	24	10	29	9	21	18	13	16	S
S7	27	23	28	10	28	9	27	15	10	20	M37
S8	31	22	36	10	29	14	27	9	28	17	M31a1
S9	33	22	30	10	16	15	31	22	10	17	M
S10	29	22	28	10	26	12	27	17	10	19	N
S11	27	22	14	10	28	18	28	19	11	21	S
S12	31	22	28	14	17	9	34	15	12	20	M
S13	32	22	28	10	28	14	27	14	11	26	M33a
S14	32	26	15	10	26	14	27	16	12	16	S
S15	29	22	27	14	30	17	27	17	14	18	N
S16	29	22	14	10	23	14	27	7	12	17	N
S17	30	19	19	10	23	9	27	17	4	19	M3
S18	28	22	28	12	23	9	27	11	10	21	U8
S19	35	22	18	10	17	17	28	17	13	19	F1
S20	29	22	27	10	28	13	27	18	10	17	F1
S21	30	22	19	10	28	12	27	14	11	17	D4
S22	27	26	28	10	26	12	31	14	12	28	M3
S23	29	26	28	10	28	12	27	11	12	18	U7
S24	33	22	28	10	28	14	34	13	4	16	H
S25	27	22	24	11	23	15	27	18	25	19	M24
S26	28	22	24	10	30	16	29	13	24	22	D4
S27	29	22	22	10	29	14	27	15	14	17	M33a
S28	27	26	24	10	24	18	28	17	14	19	M30
S29	28	22	24	10	26	12	28	18	15	20	H
S30	27	22	19	10	30	17	28	14	11	21	N
S31	28	22	26	10	32	9	28	9	21	19	F1
S32	29	22	24	10	23	12	27	11	19	19	M33a
S33	29	22	35	10	32	14	28	9	13	18	M3
S34	29	22	29	10	29	9	22	17	11	20	M34a
S35	29	25	30	10	29	12	30	17	13	17	G2
S36	27	22	27	10	28	17	27	17	12	21	G2
S37	35	22	27	10	23	16	28	11	10	17	M7a
S38	31	22	28	10	28	9	27	14	4	18	B4
S39	27	22	31	10	27	9	30	18	13	19	Z3a
S40	27	22	24	10	25	15	34	17	12	20	U7
S41	32	22	19	10	28	17	28	18	17	22	M10
S42	32	22	19	10	27	14	27	15	12	19	M37
S43	36	22	28	10	27	14	27	11	11	17	D4
S44	29	22	20	10	27	12	30	18	12	17	M9
S45	27	22	28	10	17	13	34	16	13	19	M5d
S46	30	25	22	10	23	12	27	11	15	22	M33a
S47	29	22	24	13	30	10	27	16	13	19	M33a
S48	28	22	28	10	24	12	34	15	16	18	M33a
S49	32	22	28	10	27	12	32	11	16	19	D4
S50	28	22	28	10	26	17	28	14	11	20	F1
S51	30	22	27	10	25	14	27	15	18	18	U7
S52	33	22	27	11	23	14	28	21	14	19	M34a
S53	29	22	28	10	22	17	27	17	12	27	A
S54	28	22	28	10	27	16	27	14	10	18	F1
S55	30	22	19	10	23	9	27	18	11	19	M3

4.4.7 Geographical Region wise gene pool contribution of Maharjan population

The Maharjan population was found with 25 different haplogroups and sub-haplogroups. Among different haplogroups they have been descendants of different macrohaplogroups arose by the M and N root of African haplogroup L3. Hence, the present Maharjan population was found to be contributed by different gene pools. In this research work gene pool contribution has been studied in terms of geographic region and major ethnic population of the worlds. Major haplogroup M that included M3, M5, M33, M30, M37, M31, M34 of South Asian origin were found in 57.56% of Maharjan population. Macrohaplogroup M is descendants from South Asia and covers about 70% of Indian mtDNA lineage (Chandrasekar et al., 2009). East Asian haplogroups such as F1, G, D4 were prevalent in 19.99% of Maharjan population and are said to be originated in Ainu, Japanese, Mongol, Tibetan, Yunnan, Nicobar islands, Arunanchal pradesh (Derenko et al., 2008). Similarly, 18.85% of Maharjan population under study showed haplogroup confined to Western Eurasian that include Iran, Turkey, Croatia, Denmark, Greece etc. (Geyden et al., 2009). Haplogroup A and Z which was originated in Central Asia was observed with frequency 1.8% each and thus Central Asian gene pool in Maharjan population was found to be 3.6% only suggesting Maharjan population Gene pool contribution from Central Asia is low compared to South Asian, East Asian and Western Eurasian. Thus, region wise genepool contribution in Maharjan population was dominated by South Asia followed by East Asia, Western Eurasian and Central Asia.

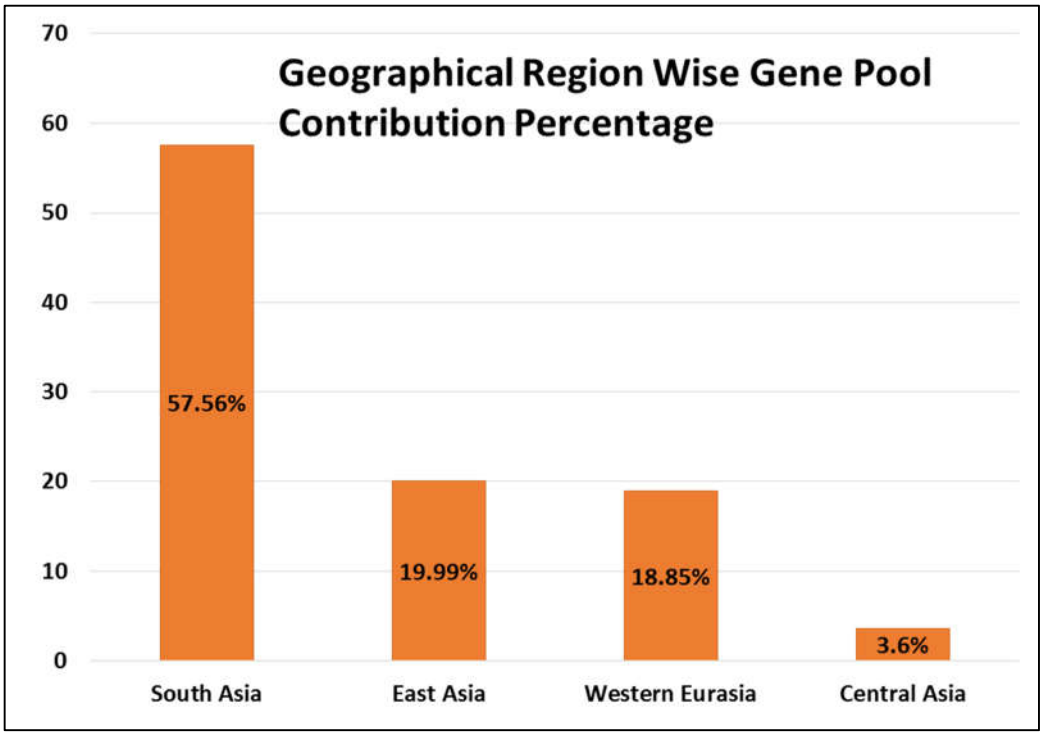


Fig. 17: Gene pool contribution of different geographical region to the Maharjan population

CHAPTER 5

DISCUSSION

This study was carried out to assess the triplet repeat distribution among the unrelated healthy individuals belonging to Maharjan cast of Newar ethnic group residing in Kathmandu. Another part of this study was to determine the mitochondrial haplogroup by sequencing of control region of mitochondrial genome.

5.1 Trinucleotide Repeat Length Distribution

Trinucleotide repeat expansion are associated with amplification of CAG, CGG or GAA repeats contained within specific genes. Polymorphic forms of these repeats occur in general population and when it reaches beyond a threshold size they become pathogenic. The prevalence of neurodegenerative diseases caused by such trinucleotide repeat expansions are restricted to a few studies from isolated geographical regions and do not reflect the real occurrence of the disease. Normal and disease ranges for triplet repeat disorders are reported to vary considerably between populations. Screening of populations for such polymorphisms helps to establish the normal and expanded ranges for that particular geographical region, which will enable proper molecular diagnosis. Moreover, repeats which are large but still within the normal range, referred to as Large Normal (LN) alleles, are known to be indicators of disease prevalence (Alluri et al., 2007).

Normal and LN allele distributions were assessed from the healthy individuals of Maharjan population (n = 55). Capillary electrophoresis (CE) was used for DNA fragment analysis to determine trinucleotide repeat distribution of 10 different triplet disorder in 55 control individuals.

Trinucleotide repeat length distribution was studied at various loci including SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DRPLA, FXTAS and DM1 in an ethnic population of Nepal. Study was aimed to examine whether there was any predisposition of particular type of tandem repeat disorder or not. The PCR conditions resulted in consistent amplification of the normal-size alleles were analyzed, which were observed as well-defined peaks with no background. No expanded alleles were found at any studied locus. However, presence of large normal alleles from which further expansion of the disease range takes place were observed in SCA3 (above 27 repeats), SCA7 (above 12 repeats), and in DRPLA (above 17 repeats).

SCA1 locus allele size ranged from 22 to 36 (CAG)n repeats in our studied population. The 36-repeat allele in the SCA1 locus is borderline, in some populations being an intermediate allele with reduced penetrance (Zühlke et al., 2002). Normal size of allele in SCA1 locus in our population is comparable with the whites. At the SCA1 locus frequency of large normal alleles greater than 30 and 31 repeats is 0.1 in among 110 chromosomes studied. Our data found no intermediate allele for SCA1. SCA1 is significantly higher

among white than in Japanese. Size of normal alleles at SCA1 loci in Japanese population & Caucasian population were observed in a range between 17-38 & 25-37 respectively (Basu et al., 2000). Normal alleles of SCA1 locus of our population was comparable with the normal alleles of both Japanese and Caucasians. Among SCA subtypes in Indian population SCA1 represents a larger proportion. Normal range of SCA1 repeats in North Indian population was observed in range 7 – 37 (Saleem et al., 2000).

A 22-repeat allele is described in other populations as the most frequent at SCA2 locus. The frequency of alleles of more than 22 repeats is higher in white, Indian and Japanese population (Freund et al., 2009). In our study, allele size for SCA2 locus ranged from 19 to 26 (CAG)n with 22 repeat as the most frequent repeat. About 72.7% of 110 chromosomes constituted of 22 repeat. Large normal allele greater than 22 repeats in our population was 7.2%. So, in our studied population out of 110 chromosomes chances for expansion of allele is about 7.2%. SCA2 is more prevalent in white population and Indian population. Normal CAG repeats range for SCA2 in North Indian, Koreans, Japanese and Caucasian are 18−30, 19−27, 15−32, 22−29 respectively (Saleem et al., 2000; Kim et al., 2001; Takano et al., 1998). The range of CAG repeat for SCA2 in our studied population is comparable to Korean population.

Among the different locus studied large normal allele (>27 repeats) of SCA3 was comparatively higher. Frequency of large normal allele varies from population to population. In our study, we analysed CAG repeat number among 55 normal individual at SCA3 locus which ranged from 14 to 36 repeats with 28 repeat being the most frequent in the pooled set. Large normal allele at SCA3 locus in our population constituted of 27.7%. Frequency of large normal alleles at SCA3 locus was comparatively higher than other repeat disorder so there might be high incidence of SCA3 patients in the studied population. Alleles with a number of (CAG)n greater than 27 are observed in the Portuguese, American, Japanese and African populations, of which there is a high incidence in SCA3 patients. Prevalence of SCA3 is reported to be higher in Japanese populations compared to Indian populations, although frequency of alleles greater than 27 repeats at SCA3 locus in Japanese population was less than that of eastern population of India. Also in case of Caucasians population prevalence of SCA3 varies considerably. Molecular analysis of unrelated Brazilian families with SCA3 showed that normal alleles ranged from 12 to 33 (CAG)n (Chattopadhyay et al., 2003). Normal range of CAG repeats of SCA3 in our study population is comparable with North Indian population with repeat range 14-37 (Saleem et al., 2000).

In another study of DRPLA, normal allele ranged from 6-35 repeats in Japanese population and 8-21 repeats in Caucasians. DRPLA is less frequent in white population than Japanese population. In our study we detected no expansion in DRPLA loci among 55 samples. Range of normal allele at DRPLA loci was found to be 7-22 (CAG)n repeats which was similar in range with the Caucasians. Large normal alleles greater than 17 repeat constituted of about 11.8 %.

The distribution of Normal CAG repeat in SCA7 Loci in 110 chromosomes of 55 healthy individuals of Newar population ranged from 6-14 with 10 CAG repeat being the most common allele observed. The 10 CAG allele represented 80% of the alleles among normal Newar population of study. Large normal alleles in SCA7 are considered repeats greater than 12, which was only 2.7% of the total alleles analysed in this study. No intermediate alleles were found. In a study carried out in by Garcia-Velazquez et al., 2013 in Mexican population, CAG repeats in normal allele ranged from 8 to 13 with 10 CAG repeat being the most frequent. CAG repeat distributions at the SCA7 loci in North Indian population was observed in range 9-14 (Saleem et al., 2000). Our study also revealed comparable range of CAG repeats in SCA7 loci with North Indians. The SCA7 mutation has been identified in various ethnic groups and geographical regions around the world but the prevalence of SCA7 has been shown to be the highest among cases of familial ADCAs in South Africa (22%) (Johansson et al., 1998).

CAG repeat length in SCA12 loci ranged from 9 -18 repeats with 9 CAG repeats being the most frequent. There were no intermediate allele and large normal allele in our studied population. Since, worldwide incidence of SCA12 is low, it can be possible that occurrence of predisposed CAG intermediate allele in SCA12 loci in our studied population is low. In case of SCA8 our study determined CTG repeat ranges from 11 to 32 with 23 repeats being the most frequent. SCA8 are not so common worldwide. Juvonen et al., 2005 carried out a study among Finnish population that revealed CTG repeat range from 15 to 35. Range of CTG repeat is comparable with our study. Repeat greater than 35 are prone to expansion of repeats and contracting a disease.

In another polyglutamine disorder HD, allele size at HD locus ranged from 12 to 28 (CAG)n with 17 repeat allele being the most frequent. Autosomal dominant disorder HD have average CAG tract length in general population from 16 to 29 repeats. CAG repeat size distributions in HD ranged between 11 and 31 repeats with 16 repeat allele was the most frequent in 12 ethnic populations of India (Pramanik et al., 2000).

For non-polyglutamine disorders, FXTAS and DM1 locus were examined for CGG & CTG repeats respectively. FXTAS allele size ranged from 18 – 34 CGG repeats whereas DM1 allele size ranged from 1–28 CTG repeats. FXTAS is a late-onset neurodegenerative disorder which arise from full mutation (>200 CGG repeats in the gene *FMR)* or premutation (55 to 200 CGG repeats) alleles. Normal alleles carry approximately 5-44 CGG repeats, highest percentage of individuals have approximately 29-31 repeats and smaller. Permutation *FMR1* alleles are transmitted into next generations resulting in an affected individual only when a full mutation is produced (Hagerman et al., 2015).

Analysis of 110 alleles showed 19 distinct allele sizes with different CTG lengths of *DMPK* ranging from 1-28 repeats in 55 normal individuals from Maharjan population. 11 and 12 was most frequent number of CTG repeats (17.2%). >19 CTG repeats in *DMPK* are considered as large normal alleles that helps to correlate prevalence of DM1 (Kwon et al., 2010). Frequency of larger normal alleles in our samples was 3.6%. In most of the

populations the range of CTG repeats in *DMPK* have at least 5 repeats however, our samples showed one and four CTG repeats which might me due to any errors during fragment analysis. CTG repeats in Serbian, Iranian and African-American population has shown range of 5-28 CTG repeats in their healthy individuals whereas Korean and European population showed range of 5-36 CTG repeats (Imbert et al., 1993; Krndija et al., 2005).

Triplet repeat length distribution analysis in an ethnic Nepali populations showed occurrence of normal alleles. 10 different trinucleotide repeats disorder length expansion were examined in 55 normal individuals of same ethnic background and it revealed all normal alleles. However, there were presence of large normal alleles that causes further expansion. SCA3 had highest percentage of large normal alleles compared to other disorder examined and the range of repeats were comparable with normal individuals of Indian population.

5.2 Identification of mtDNA Haplogroup:

Mitochondrial DNA undergoes uniparental inheritance so, mtDNA can be used to trace maternal lineage of different population residing in different geographical region. mtDNA haplogroup was identified by sequencing the HVS I and HVS II of mitochondrial D-Loop and by comparing with rCRS (Cambridge Reference Sequence). On the basis of variations in the D-loop region, phylogenetic tree can be created which describes the mtDNA hapolgroup. Newars are the sixth largest ethnic groups of Nepal. Newars are the original inhabitants of the Kathmandu valley, but their origins are shrouded in mystery. They speak Nepal Bhasa, a Tibeto- Burmease language, which indicates potential origin in the east, however, their physical features range from distinctively Mongoloid to Indo-Aryan (Gellner, 1986).

5.2.1 Mitochondrial Haplogroup Diversity:

Mitochondrial DNA is characterized by a high mutation rate and these mutations form groups of stable haplotypes and are known as haplogroups. Mitochondrial DNA haplogroups tend to be geographically restricted and they are used to genetically distinguish populations. Displacement loop (D-loop) is most variable part of mtDNA. It has most polymorphics sequences and is objects of many studies and researches of the roots of populations and human evolution (Nesheva, 2014).

In total 5 major haplogroups M, N, U, H, R were found among the 55 healthy individual from maharjan population. The presence of Asian specific haplogroups like M and N were revealed in this study. Among these haplogroup M was most common with 58.2% of the total studied population. In our studied population haplogroup M with its sub-haplogroups identified are M5, M33, M12'G, M4'67, M3, M24'41, M10, M62'68, M9, M34'57, M7, M8, M31. The mitochondrial haplogroup M which was first regarded as an ancient marker of East-Asian origin found at high frequency in India. M2, M3, M5, M9, M30, M33, M34, M35, M38 sub-haplogroup of M were originated in South Asia from

60,000 YBP. They contribute more than 60% of South Asian Maternal lineage. Macro Haplogroup M covers more than 70% of Indian mtDNA lineages (Chandrasekar et al., 2009). In a study carried out by Fornarino et al, 2009, a great majority of the Tharu mtDNA (largest indigenous population in Terai region of Nepal) is represented by lineages shared or derived from Indian haplogroup M (M31, M33, M35, M38). Sub clades M33 and M31 were also seen in our study population suggesting that this Newar population under study share some common maternal ancestry with the Indian population. Bhandari et al in 2015 has reported M9a haplogroup & its subclades distributed among Sherpa popon which is highly prevalent among Tibetans suggesting close maternal relationship between them. However, haplogroup M9a was not yielded in any of our studied samples suggesting no any maternal relationship with Tibetans.

Macrohaplogroup N considered as second haplogroup that have diverged from African lineage L3, was found with frequency of 18.2% in this study population of Maharjan community. Haplogroup N is the ancestor of many haplogroups found in Europe, Middle East, and Asia. The origin of this lineage occurred soon after or possibly even during the migration out of Africa some 39,000–51,000 YBP, and is typically considered a southwest Eurasian lineage (Stewart & Chinnery, 2015). Haplogroups A, F1 and R are descendent from macrohaplogroup N. These haplogroups are most prevalent in Central Asian and Western Eurasian. Haplogroup A was orginated in Central Asian 50,000 YBP was most frequent in Tibetian (Derenko et al., 2007). In our study population of Maharjan community only 1.8% of represented this haplogroup. Suggesting that Tibetian ancestry is less likely in this population. Similarly, the frequency of haplogroup F in our study population was 9.09%. The haplogroup F is believed to be originated in East Asia descendent from haplogroup R9 with time of origin 43,000 YBP (Soares et al., 2009).

Haplogroup Z was also identified in our study population. The frequency of this haplogroup was 1.8% with Z3 as sub-haplogroup. Haplogroup Z originated in Central Asia and distributed throughout East Asia. Haplogroup D4 that is East Asia specific was also observed in our study population with frequency 7.3%.

The haplogroup G which was descendant from M12, was originated 35,000 YBP (Soares et al., 2009) in East Asia. This haplogroup was observed in our Maharjan population with frequency 3.6%. In a study by Fornarino et al., 2009 this haplogroup was reported in Tharu of Chitwan with frequency 23.3% and Tharu of Morang with 12.5%.

In India most frequent sub-clade of R is haplogroup U. Haplogroup U represents overlap between western-Eurasian and Indian mtDNA lineages which is estimated age of 51,000-67,000years. About 7.3 % of population under study was observed with haplogroup U. U7 and U8 were identified as the subset of Haplogroup U. Haplogroup U7 was found in our study population with frequency 5.5% whereas, U8 with 1.8%. U7 is present commonly in Gujarat of India and distributed from India to Iran (Palanichamy et al, 2004).

Study of Mitochondrial DNA haplogroup gives an idea about the maternal lineage of the Newar population. Phylogenetic tree showed variations in genetic structure within the sample population. After assigning each haplogroup for mitochondrial control region it was found that the largest contribution was from South Asia (57.56%) followed by East Asia (19.99%), Western Eurasia (18.85%) and Central Asia (3.6%) and language but there is still high genetic diversity among them.

Principal component analysis (PCA) revealed that Maharjan population showed genetic affinity between Altaian Kazakhs, Uzbeks (Derenko et al., 2008), Tharu of Eastern Terai (Fornarino et al., 2009) and Kathmandu population (Gayden et al., 2013) because these population shares common haplogroups. However Kathmandu population doesnot reveal which ethnic group/caste was taken in the study.

5.3 Comparison of Different TNRs with mtDNA Haplogroup of Maharjan Sub-ethnic Group

There were very few samples sharing same mtDNA haplogroups. mtDNA haplogroup identification in Maharjan population through sequencing of D-loop revealed high genetic diversity within the population. In order to identify the pathogenic role of the mitochondrial genome it requires more extensive surveys of the mtDNA sequences in different populations and patient groups. Although few samples shared same haplogroups comparision of repeat length of 10 different trinucleotide repeat disorder loci showed no correlation among them. Repeat length of 10 different trinucleotide disorder was variable among the samples. However, trinucleotide repeat of SCA2, SCA7 and FXTAS showed common repeat of 22, 10 and 27 respectively with high frequency but there was no correlation with the mtDNA haplogroup.

Also, comparing the trinucleotide repeat length and their corresponding haplogroup in each individual from two different locations also did not show any correlation. Maharjan population from two different locations i.e. neither Kirtipur nor Harisidhhi showed any correlation with regard to mtDNA haplogroups along with trinucleotide repeat length.

CHAPTER 6

SUMMARY

Under this study, 55 unrelated healthy Maharjan individuals who share common ethnic background were investigated for trinucleotide repeat length distribution of 10 different repeat disorders viz. SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DM1, FXTAS and DRPLA. SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, DRPLA are subtypes of Spinocerebellar ataxias (SCAs) caused by CAG repeat expansions along with HD. While FXTAS and DM1 are caused by CGG and CTG repeat expansion. Trinucleotide repeats occur as polymorphic forms in general population. However, beyond a threshold size they become pathogenic. Therefore, screening of populations for such polymorphisms helps to establish the normal and expanded ranges for that particular geographical region, which will enable proper molecular diagnosis. Repeats which are large but still within the normal range, referred to as Large Normal (LN) alleles, are known to be indicators of disease prevalence. Therefore to elucidate the normal repeat range and LN repeat frequencies variations in number of CAG/CGG/CTG repeats were analysed. For molecular analysis of 55 individuals belonging to Maharjan population DNA was isolated by salting out method. To analyse the repeat length each loci of SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DM1, FXTAS and DRPLA were amplified with specific primer at the region of interest and PCR products were subjected to capillary electrophoresis for fragment analysis. Fragment analysis showed chromatogram with two peaks for two alleles of respective gene. All the repeat length were found to be within normal range. Repeat range for SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DM1, FXTAS and DRPLA observed in our population are 22-36, 19-26, 14-36, 6-14, 11-32, 9-18,12-28, 1-28, 18-34 and 7-22 respectively. All the repeat range were within normal range when compared with data of various other populations. However, large normal alleles that are considered as the reservoirs for the generation of new expanded alleles were also examined. SCA1, SCA2, SCA3, SCA7, DRPLA constituted of large normal alleles with frequency 10%, 7.2%, 27.3%, 2.7% and 11.8% respectively. High frequencies of LN alleles are considered to be indicators of disease prevalence. This result illustrated that in frequency of LN alleles at the SCA3 locus was higher (27.3%) in Maharjan population under study. There was no correlation of trinucleotide repeat length among the Maharjan individuals from same location i.e. neither Kirtipur nor Harisiddhi was observed. Since, study of large normal allele is indirect measure of prevalence of disease, the result shows that incidence of SCA3 could be higher in studied population among other trinucleotide repeat disorder.

Apart from the investigation of trinucleotide repeat length distribution this study was also carried out to identify the mtDNA haplogroup. Mitochondrial lineage was determined to know the genetic distribution of studied population. Mitochondrial D-loop is the noncoding region which contains hypervariable regions that provides insight about the genetic structure of the population on the basis of the variations. D-loop

includes all the base pairs from nucleotide position 16024-16569, 1-576. For identififcation of Mitochondrial Haplogroup PCR amplification of D-loop region was done using 3 different primers that covered approximately 1122bp of D-loop region. PCR amplification was checked on 2% agarose gel to ensure its correct amplification. Then sequencing of these amplified D-loop region was done. Editing, alignment for D-loop sequence were done by using rCRS as a reference sequence on DNA star software. Each variations were noted and haplogroup was obtained by using software HaploGrep. With the help of observed haplogroups in this study and various haplogroups from published and unpublished data the principal component analysis (PCA) was done to get insights about genetic affinity using software SPSS var 16.1. The most frequent mtDNA haplogroup was found to be haplogroup M (58.2%) followed by N, H, R and U with frequency 18.2%, 3.6%, 12.7% and 7.3% respectively. Region-wise gene pool was also calculated and highest geographical gene pool contribution was from South Asia (57.56%) followed by East Asia (19.99%), Western Eurasian (18.85%) and Central Asia (3.6%). Central Asia was least contributed gene pool in this population. On the other hand PCA analysis showed diverse affinity by showing its relatively closeness with East Asian population (Altai, Uzbeck and Mongolian), South Asian population (South India, North India) and Population of Nepal (Kathmandu, Nepal_Mix, Nepal_Kat, Tharu E). Since, the Maharjan population in PCA analysis lies in mid-way it shows that this population shares its haplogroup with all other population taken under comparision. It reveals that there was higher level of diversity within the population.

After determination of variation in trinucleotide repeat numbers in 10 different disorders and identification of mitochondrial haplogroup, correlation between repeat number and similar haplogroups were investigated. There was no correlation between occurrence of trinucleotide repeat number and mitochondrial haplogroup. This is because of great diversity of mitochondrial haplogroup in Maharjan population and lack of data from clinically diagnosed positive samples, the significant contribution could not be observed between TNR length distribution in SCAs, HD, FXTAS, DM1 with that of haplogroup distribution of Maharjan population.

CHAPTER 7

CONCLUSION

Normal and disease ranges for trinucleotide repeat disorders are reported to vary considerably between populations. So, it is necessary to screen populations for such polymorphisms to establish the normal and expanded ranges for that particular geographical region which will enable variant identification and standardization for proper molecular diagnosis. Variation in TNR in SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DM1, FXTAS and DRPLA was identified by amplifying specific region of interest in respective loci and subjected to capillary electrophoresis for fragment analysis. The statistical data from our study allowed us to identify the alleles and their frequencies of TNR numbers in 10 different trinucleotide repeat disorders. Repeat range for SCA1, SCA2, SCA3, SCA7, SCA8, SCA12, HD, DM1, FXTAS and DRPLA observed in our population are 22-36, 19-26, 14-36, 6-14, 11-32, 9-18,12-28, 1-28, 18-34 and 7-22 respectively. Population genetic analysis allowed the normal trinucleotide repeat to be defined. The normal range of the repeat size allows us to distinguish between unaffected and affected individuals for diagnostic purpose. To understand the origins of the mutations at these loci it is important to consider the information provided by the normal range variation so as to understand the dynamics of these loci. Repeats which are large but still within the normal range, referred to as large normal (LN) alleles, are known to be indicators of disease prevalence and the percentage of LN alleles would give an indirect reflection of the prevalence of the disease in a given population. Maharjan population under study showed all trinucleotide repeats within normal range. However, LN alleles of SCA3 was comparatively higher (27.2%) than other trinucleotide repeat disorders studied indicating incidence of SCA3 in Maharjan population.

Study of Mitochondrial haplogroup helped us to identify the mitochondrial haplogroup composition. Mitochondrial D-loop is the noncoding region which contains hypervariable regions that provides insight about the genetic structure of the population on the basis of the variations. Sequencing of mitochondrial D-loop in 55 healthy individuals of Maharjan population revealed that South Asia (57.56%) serves as highest contributor of maternal genetic component followed by East Asia (19.99%), Western Eurasian (18.85%) and Central Asia (3.6%). However, the mtDNA haplogroup also indicated that there is high genetic diversity within the studied population. The reason for high genetic diversity within same ethnic population might be due to accumulation of more mutation and gene flow from other population. This result also showed that there is no correlation between trinucleotide repeat expansion and particular mtDNA haplogroup, however, the present data will contribute for the frequency distribution of TNR related to SCAs, HD, FXTAS and DM1. Moreover, Maharjan individuals from same location i.e. neither Kirtipur nor Harisiddhi showed any correlation with trinucleotide repeat length with that of their mtDNA haplogroup.

RECOMMENDATIONS

1. For the proper molecular diagnosis of expanded alleles in trinucleotide disorder study on the percentage of LN alleles in large sample size of various ethnic population can be done to identify prevalence of the disease in that given population.
2. For identification of disease prevalence of trinucleotide disorder age specific research can be done for classification of repeats according to age.
3. Clinically diagnosed positive cases of each trinucleotide disorder and its frequency distribution should also be studied to identify minimum and maximum repeats of expanded alleles in given population.
4. For study of maternal lineage whole mtDNA sequencing can be done to get more reliable results.
5. Genome wide scan should be done which gives detailed genetic structure and diversity of that particular population.

REFERENCES

Abdi H., and William L.J., (2010) “Principle component analysis”. Wiley interdisciplinary Reviews: Computational Statistics, 2: 433 459. Doi:10.1002/wics.101

Alluri R.V., Komandur S., Wagheray A., Chaudhuri J.R., Meena A.K., Jabeen A., Chawda K., Subhash K., Krishnaveni A. and Hasan Q. (2007) “Molecular Analysis of CAG Repeats at Five Different Spinocerebellar Ataxia Ioci: Correlation and Alternative Explanations for Disease Pathogenesis” Molecules and cells. Dec 31; 24(3):338.

Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M. and Howell N. (1999) “Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature genetics, 23 (2), 147-147

Bates G., Harper P. and Jones L. eds (2002) “Huntington's Disease. Oxford Monographs on Medical Genetics” Oxford University Press.

Bhandari S., Zhang X. & Cui C. (2015) “Genetic evidence of a recent Tibetan ancestry to Sherpas in the Himalayan region” Scientific reports, 5.

Budworth H. and Mcmurray T.C. (2013) “A brief history of triplet Repeat Diseases” NIH public access; 1010:3 -17

Calloway C.D., Duewer D.L., Redman J.W., Butler J.M., Kline M.C. and Vallone P.M., (2005) “Mitochondrial DNA typing screens with control region and coding region SNPs” Journal of Forensic Science, 50(2), pp.JFS2004293-9.

Chandrasekar A., Kumar S., Sreenath J., Sarkar B.N., Urade B.P., Mallick S., Bandopadhyay S.S., Barua P., Barik S.S., Basu D. and Kiran U., (2009) “Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian corridor” PloS one, *4*(10), p.e7447.

Charles P., Camuzat A., Benammar N., Sellal F. Destée A., Bonnet A.M., Lesage S., Le Ber I., Stevanin G., Dürr A. and Brice A. (2007) “French Parkinson's Disease Genetic Study Group. Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism?” Neurology; 69:1970–5

Cholfin J.A., Sobrido M.J., Perlman S., Pulst S.M. and Geschwind D.H., (2001) “The SCA12 mutation as a rare cause of spinocerebellar ataxia” Archives of neurology, *58*(11), pp.1833-1835.

Chung M.Y., Ranum L.P., Duvick L.A., Servadio A., Zoghbi H.Y. and Orr H.T., (1993) “Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I” Nature genetics, *5*(3), pp.254-258.

Dryland P.A., Doherty E., Love J.M. and Love D.R., (2013) “Simple repeat-primed PCR analysis of the myotonic dystrophy type 1 gene in a clinical diagnostics environment” Journal of neurodegenerative diseases.

Durr A., Stevanin G., Cancel G., Duyckaerts C., Abbas N., Didierjean O., Chneiweiss H., Benomar A., Lyon-Caen O., Julien J. and Serdaru M., (1996)“Spinocerebellar ataxia 3 and

Machado-Joseph disease: clinical, molecular, and neuropathological features" Annals of neurology, *39*(4), pp.490-499.

Dúshláine C.T.Ó., (2006) "Bioinformatic detection and analysis of functionally important genetic variation".

Faruq M., Srivastava A.K., Singh S., Gupta R., Dada T., Garg A. & Mukerji M. (2015). "Spinocerebellar ataxia 7 (SCA7) in Indian population: predilection of ATXN7-CAG expansion mutation in an ethnic population" The Indian journal of medical research, 141(2), 187.

Fornarino S., Pala M., Battaglia V., Maranta R., Achilli A., Modiano G., & Santachiara-Benerecetti S. A. (2009) "Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation" BMC Evolutionary Biology, 9(1)

Freund A.A., Scola R.H., Teive H.A., Arndt R.C., Costa-Ribeiro M.C.V.D., Alle L.F. and Werneck L.C., (2009) "Spinocerebellar ataxias: microsatellite and allele frequency in unaffected and affected individuals" Arquivos de neuro-psiquiatria, 67(4), pp.1124-1132.

García-Velázquez L.E., Canizales-Quinteros S., Romero-Hidalgo S., Ochoa-Morales A., Martínez-Ruano L., Márquez-Luna C., Acuña-Alonzo V., Villarreal-Molina M.T., Alonso-Vilatela M.E. and Yescas-Gómez P. (2014) "Founder effect and ancestral origin of the spinocerebellar ataxia type 7 (SCA7) mutation in Mexican families" Neurogenetics, 15(1), pp.13-17.

Gatchel J. R., & Zoghbi H. Y. (2005) "Diseases of unstable repeat expansion: mechanisms and common principles" Nature Reviews Genetics, 6(10), 743-755.

Gellner D.N. (1986) "Language, caste, religion and territory: Newar identity ancient and modern". European Journal of Sociology/Archives Européennes de Sociologie, 27(1), 102-148.

Gibbs R.A., Belmont J.W., Hardenbol P., Willis T.D., Yu F., Yang H., Ch'ang L.Y., Huang W., Liu B., Shen Y. and Tam P.K.H., (2003) "The international HapMap project" Nature, *426*(6968), pp.789-796.

Gispert S., Twells R., Orozco G., Brice A., Weber J., Heredero L., & Hillermann, R. (1993). "Chromosomal assignment of the second locus for autosomal dominant cerebellar ataxia (SCA2) to chromosome 12q23–24.1" Nature genetics, 4(3), 295-299.

Hagerman P.J. and Hagerman R.J. (2015) "Fragile X–associated tremor/ataxia syndrome. Annals of the New York Academy of Sciences" Mar 1; 1338(1):58-70.

Hattori M., Yuasa H., Takada K., Yamada T., Yamada K., Kamimoto K. & Uchida M. (1999) "Genetic analysis of a dentatorubral-pallidoluysian atrophy family: relevance to apparent sporadic cases" Internal medicine, 38(3), 287-289.

Hegde M.V. and Saraph A.A., (2011) "Unstable genes unstable mind: beyond the central dogma of molecular biology" Med Hypotheses, 77:165–170. [PubMed: 21507580]

Herrnstadt C., Elson J.L., Fahy E., Preston G., Turnbull D.M., Anderson C., & Howell N. (2002) "Reduced-median-network analysis of complete mitochondrial DNA coding-

region sequences for the major African, Asian, and European haplogroups" The American Journal of Human Genetics, 70(5), 1152-1171.

Huynh D.P., Yang H.T., Vakharia H., Nguyen D., & Pulst S. M. (2003) "Expansion of the polyQ repeat in ataxin-2 alters its Golgi localization, disrupts the Golgi complex and causes cell death" Human molecular genetics, 12(13), 1485-1496.

Imbert G., Kretz C., Johnson K., Mandel J.L. (1993) "Origin of the expansion mutation in myotonic dystrophy" Nature Genetics; **4**:72–76.

Ingman M., Kaessmann H., PaÈaÈbo S. and Gyllensten U., (2000) "Mitochondrial genome variation and the origin of modern humans" Nature, 408(6813), pp.708-713.

Johansson J., Forsgren L., Sandgren O., Brice A., Holmgren G., & Holmberg M. (1998) "Expanded CAG repeats in Swedish spinocerebellar ataxia type 7 (SCA7) patients: effect of CAG repeat length on the clinical manifestation" Human molecular genetics, 7(2), 171-176.

Juvonen V., Hietala M., Kairisto V. & Savontaus M. L. (2005) "The occurrence of dominant spinocerebellar ataxias among 251 Finnish ataxia patients and the role of predisposing large normal alleles in a genetically isolated population" Acta neurologica scandinavica, 111(3), 154-162.

Karmin M. (2005) "Human mitochondrial DNA haplogroup R in India: dissecting the phylogenetic tree of South Asian-specific lineages*"* (Doctoral dissertation, MSc. thesis, University of Tartu, Estonia. [Unpublished]).

Kim J.Y., Park S.S., Joo S.I., Kim J.M. and Jeon B.S. (2001) "Molecular analysis of spinocerebellar ataxias in Koreans: frequencies and reference ranges of SCA1, SCA2, SCA3, SCA6, and SCA7" Mol. Cells 12, 336−341.

Kivisild T. (2015) "Maternal ancestry and population history from whole mitochondrial genomes. Investigative genetics" Investigative Genetics. 6(1), p.1.

Kivisild T., Bamshad M.J., Kaldma K., Metspalu M., Metspalu E., Reidla M., Laos S., Parik J., Watkins W.S., Dixon M.E. and Papiha S.S., (1999) "Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages" Current Biology, 9(22), pp.1331-1334.

Kivisild T., Kaldma K., Metspalu M., Parik J., Papiha S. and Villems R., (1999) "The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World" In Genomic diversity (pp. 135-152). Springer US.

Kivisild T., Tolk H.V., Parik J., Wang Y., Papiha S.S., Bandelt H.J. and Villems R., (2002) "The emerging limbs and twigs of the East Asian mtDNA tree" Molecular Biology and evolution, 19(10), pp.1737-1751.

Kloosterman A.D. (2003) "Current and Future Developments in Forensic DNA Typing" Scriptie Biomedische Wetenschappen van E. Besselink

Krndija D., Savic D., Mladenovic J., Rakocevic-Stojanovic V., Apostolski S., Todorovic S. and Romac S., (2005) "Haplotypeanalysis of the DM1 locus in the Serbian population" Acta neurologica scandinavica; 111:274-277.

Kwon M.J., Lee S.T., Kim B.J., Sung D.H., Kim J.W. & Ki C.S. (2010) “Haplotype analysis of the myotonic dystrophy type 1 (DM1) locus in the Korean population” Annals of Clinical & Laboratory Science, 40(2), 156-162.

La Spada A.R. and Taylor J.P. (2010) “Repeat expansion disease: progress and puzzles in disease pathogenesis” Nature Review Genetics, 11:247–258. [PubMed: 20177426]

Lastres-Becker I., Rüb U. and Auburger G., (2008) “Spinocerebellar ataxia 2 (SCA2). The cerebellum” *7*(2), pp.115-124.

Lutz R.E. (2007) “Trinucleotide repeat disorders” In Seminars in pediatric neurology 14:26-33

McMurray C.T. (2010) “Mechanisms of trinucleotide repeat instability during human development” Nature review

Michalik A., Martin J.J. & Van Broeckhoven C. (2004) “Spinocerebellar ataxia type 7 associated with pigmentary retinal dystrophy” European journal of human genetics: EJHG, 12(1), 2.

Miller S.A., Dykes D.D. & Polesky H.F.R.N. (1988) “A simple salting out procedure for extracting DNA from human nucleated cells” Nucleic acids research, 16(3), 1215.

Mirkin S.M. (2006) “DNA structures, repeat expansions and human hereditary disorders. Current opinion in structural biology” 16(3), pp.351-358.

Mirkin S.M. (2007) “Expandable DNA repeats and Human disease” Nature Vol 447

Mishmar D., Ruiz-Pesini E., Golik P., Macaulay V., Clark A.G., Hosseini S., Brandon M., Easley K., Chen E., Brown M.D. and Sukernik R.I. (2003) “Natural selection shaped regional mtDNA variation in humans” Proceedings of the National Academy of Sciences, 100(1), pp.171-176.

Mittal U., Roy S., Jain S., Srivastava A.K. and Mukerji M. (2005) “Post-zygotic de novo trinucleotide repeat expansion at spinocerebellar ataxia type 7 locus: evidence from an Indian family” Journal of human genetics, *50*(3), pp.155-157.

Mittal U., Srivastava A. K., Jain S., Jain S. & Mukerji M. (2005) “Founder haplotype for Machado-Joseph disease in the Indian population: novel insights from history and polymorphism studies” Archives of neurology, 62(4), 637-640.

Nesheva D.V. (2014) “Aspects of ancient mitochondrial DNA analysis in different populations for understanding human evolution” Balkan journal of medical genetics. Jun; 17(1):5.

Ohshima K. & Wells R.D. (1997) “Hairpin formation during DNA synthesis primer realignment *invitro* in triplet repeat sequences from human hereditary disease genes” Journal of Biological Chemistry, 272, 16798–16806

Orr H.T. and Zoghbi H.Y. (2007) “Trinucleotide Repeat Disorders” The Annual Review of neuroscience, 30:575-621

Palanichamy M., Sun C., Agrawal, S., Bandelt, H.J., Kong, Q.P., Khan, F., Wang, C.Y., Chaudhuri, T.K., Palla, V. and Zhang, Y.P., (2004) “Phylogeny of mitochondrial DNA

macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia" The American Journal of Human Genetics, *75*(6), pp.966-978.

Pandey N., Mittal U., Srivastava A.K. and Mukerji M. (2004) "SMARCA2 and THAP11: potential candidates for polyglutamine disorders as evidenced from polymorphism and protein-folding simulation studies" Journal of human genetics, 49(11), pp.596-602.

Paulson H.L. and Fischbeck K.H. (1996) "Trinucleotide repeats in neurogenetic disorders" Annual reviews of neuroscience 19-70: 107

Pearson C.E., Tam M., Wang Y.H., Montgomery S.E., Dar A.C., Cleary J.D. and Nichol K. (2002) "Slipped-strand DNAs formed by long (CAG)·(CTG) repeats: slipped-out repeats and slip-out junctions" Nucleic acids research, *30*(20), pp.4534-4547.

Peprah E. (2012) "Fragile X syndrome: the FMR1 CGG repeat distribution among world populations" Annals of human genetics, 76(2), pp.178-191.

Pramanik S., Basu P., Gangopadhaya P.K., Sinha K.K., Jha D.K., Sinha S., Das S.K., Maity B.K., Mukherjee S.C., Roychoudhuri S. and Majumder P.P. (2000) "Analysis of CAG and CCG repeats in Huntingtin gene among HD patients and normal populations of India" European journal of human genetics: EJHG, 1;8(9):678.

Pringsheim T., Wiltshire K., Day L., Dykeman J., Steeves T. and Jette N. (2012) "The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis" Movement Disorders, *27*(9), pp.1083-1091.

Pulst S.M., Nechiporuk A., Nechiporuk T., Gispert S., Chen X.N., Lopes-Cendes I., Pearlman S., Starkman S., Orozco-Diaz G., Lunkes A. and DeJong P. (1996) "Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2" Nature genetics, 14(3), pp.269-276.

Ranum L.P. & Cooper T.A. (2006) "RNA-mediated neuromuscular disorders." Annu. Rev. Neurosci., 29, 259-277.

Rao S.M., Trividi S., Emmanuel D., Merita K. and Hynniewta M. (2010) "DNA repetitive sequences-types, distribution and function: A review" Journal of Cell and Molecular Biology 7(2) & 8(1): 1-11

Rich J., Ogryzko V.V. and Pirozhkova I.V. (2014) "Satellite DNA and related diseases" Biopolymers and Cell, 30(4), pp.249-259.

Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V. & Villems R. (2000) "Tracing European founder lineages in the Near Eastern mtDNA pool" The American Journal of Human Genetics, 67(5), 1251-1276.

Rubinsztein D.C., Leggo J., Coetzee G.A., Irvine R.A., Buckley M. & Ferguson-Smith M.A. (1995) "Sequence variation and size ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes" Human molecular genetics, 4(9), 1585-1590.

Saleem Q., Choudhry S., Mukerji M., Bashyam L., Padma M., Chakravarthy A., Maheshwari M.C., Jain S. and Brahmachari S.K. (2000) "Molecular analysis of autosomal

dominant hereditary ataxias in the Indian population: high frequency of SCA2 and evidence for a common founder mutation" Human genetics. Feb 25; 106(2):179-87

Sbisà E., Tanzariello F., Reyes A., Pesole G. and Saccone C. (1997) "Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications" *Gene*, *205*(1), pp.125-140.

Shadel G.S. and Clayton D.A. (1997) "Mitochondrial DNA maintenance in vertebrates" Annual. Rev. Biochem. 66:409

Sharma S., Saha A., Rai E., Bhat A. and Bamezai R. (2005) "Human mtDNA hypervariable regions, HVR I and II, hint at deep common maternal founder and subsequent maternal gene flow in Indian population groups" Journal of human genetics, *50*(10), pp.497-506.

Shrestha T.R., Manadhar K.D., Awasthi N.P., Thangaraj K. (2013) "Genetic structure of Sub-ethnic groups of Newar population of Kathmandu Valley" (Unpublished data).

Shrestha T.R., Manadhar K.D., Pradhan I., Thangaraj K. (2013) "Genetic Affinity of Manadhar population of Kathmandu Valley" (Unpublished data).

Srivastava A.K., Choudhry S., Gopinath M.S. Roy S., Tripathi M., Brahmachari S.K. and Jain S. (2001) "Molecular and clinical correlation in five Indian families with spinocerebellar ataxia 12" Ann Neurol 50:796–800.

Taanman J.W. (1999) "The mitochondrial genome: structure, transcription, translation and replication" Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1410(2), 103-123.

Takano H., Cancel G., Ikeuchi T., Lorenzetti D., Mawad R., Stevanin G., Didierjean O., Dürr A., Oyake M., Shimohata T. and Sasaki R. (1998) "Close associations between prevalence's ofdominantly inherited spinocerebellar ataxias with CAGrepeat expansions and frequencies of large normal CAG alleles in Japanese and Caucasian populations" Am. J. Hum. Genet. 63, 1060−1066.

Taylor R.W. and Turnbull D.M. (2005) "Mitochondrial DNA mutations in human disease" Nature Reviews Genetics, *6*(5), pp.389-402.

Thangaraj K., Chaubey G., Kividild T., Rani D.S., Singh V.K., Ismail T (2008) "Maternal footprints of Southeast Asians in North India" Hum. Hered.66, 1-9.

Todd P.K. and Paulson H.L. (2010) "RNA-mediated Neurodegeneration in repeat expansion disorders" Annals of neurology, 67(3), pp.291-300.

Usdin K. (2008) "The biological effects of simple tandem repeats: lessons from the repeat expansion diseases" Genome research, 18(7), pp.1011-1019.

Wallace D.C. (1994) "Mitochondrial DNA sequence variation in human evolution and disease" Proceedings of the National Academy of Sciences, *91*(19), pp.8739-8746.

Wang H.W., Li Y.C., Sun F., Zhao M., Mitra B., Chaudhuri T.K. & Zhang Y.P. (2012) "Revisiting the role of the Himalayas in peopling Nepal: insights from mitochondrial genomes" Journal of human genetics, 57(4), 228-234.

Whale J.W. (2012) "Mitochondrial DNA analysis of four ethnic groups of Afghanistan" Portsmouth: University of Portsmouth.

Whaley N.R., Fujioka S. and Wszolek Z.K. (2011) “Autosomal dominant cerebellar ataxia type I: a review of the phenotypic and genotypic characteristics” Orphanet journal of rare diseases, *6*(1), p.1.

Yakura H., Wakisaka A., Fujimoto S. and Itakura K. (1974) “Hereditary ataxia and HL-A” The New England journal of medicine, 291(3), pp.154-155.

Zühlke C., Dalski A., Hellenbroich Y., Bubel S., Schwinger E. and Bürk K. (2002) “Spinocerebellar ataxia type 1 (SCA1): Phenotype-genotype correlation studies in intermediate alleles” European Journal of Human Genetics, 10(3).

Websites

http:// www.omim.org/

http://Phylotree.org/tree/main.htm

http://www.haplogrep.uibk.ac.at/

http://www.dialogues-cns.com/publication/a-glossary-of-relevant-genetic-terms/

http://www.mitomap.org/

http://www.ncbi.nlm.nih.gov/

www.nepalitimes.com

https://www.ncbi.nlm.nih.gov/books/NBK1305/

Appendices

Appendix 1: Reagents used in DNA Isolation

1. RBC lysis buffer (10x)

NH4Cl	8.20 gm
NaHCO3	0.84 gm
EDTA	0.37 gm

Dissolved in 100ml of distilled water, autoclaved and stored at 4^0C. Working dilution (1x)

For 500 ml 1x RBC lysis buffer - 50ml RBC lysis buffer (10x) + 450 ml autoclaved water.

2. Nucleus Lysis Buffer (NLB)

10 mM Tris – HCL

400 mM NaCl

2 mM Na_2EDTA (pH 8.0)

Autoclaved and stored at room temperature.

For 400 ml

1M Tris HCL (pH 8.0) – 4ml

5 M NaCl– 32ml

0.5 M EDTA (pH 8.0) – 1.6ml

Final volume made up to 400 ml (Autoclaved and stored at room temperature)

3. Proteinase K solution (20 mg / ml)

20 mg of Proteinase K was dissolved in 1 ml autoclaved distilled water. Stored at 4^0C.

4. 10% SDS (sodium dodecyl acetate)

For 100 ml – 10 gm of SDS was dissolved in autoclaved distilled water. Final volume was made up to 100 ml (stored at room temperature).

5. 6M saturated NaCl solution

NaCl – 35.064 gm dissolved in distilled water and made final volume up to 100 ml.

6. TE Buffer (200ml)

10 mM Tris (pH 8.0)

1 mM EDTA (pH 8.0).

Autoclaved and stored at 4^0C

Appendix 2: PCR Protocol

SCA1

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1x
MgCl2 (25mM)	0.4 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5U/ µl
DMSO	0.5 µl	5%
MQ	4.7 µl	-
TOTAL =	**10.0 µl**	

SCA2

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1x
$MgCl_2$ (25mM)	0.4 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5U/ µl
DMSO	0.5 µl	5%
MQ	4.7 µl	-
TOTAL =	**10.0 µl**	

SCA3

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.6 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	5.0 µl	-
TOTAL =	**10.0 µl**	

SCA7

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	
PCR Buffer (5 X)	1.4 µl	0.7X
MgCl2 (25mM)	0.6 µl	1.5mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/ µl)	0.4 µl	0.4pm/ µl
R.P. (10pm/ µl)	0.4 µl	0.4pm/ µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5U/µl
DMSO	0.5 µl	5%
MQ	5.1 µl	-
TOTAL =	**10.0 µl**	

SCA8:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.8 µl	2mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	4.8 µl	-
TOTAL	**10.0 µl**	

SCA12:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.32 µl	0.8mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	4.8 µl	-
TOTAL	**10.0 µl**	

DRPLA:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.4 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	5.2 µl	-
TOTAL	**10.0 µl**	

HD:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.4 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
DMSO	0.5	5%
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	5.2 µl	-
TOTAL	**10.0 µl**	

DM1:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer (5 X)	2.0 µl	1X
$MgCl_2$ (25mM)	0.4 µl	1mM
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	5.2 µl	-
TOTAL	**10.0 µl**	

FXTAS:

Components	Working volume	Working concentration
DNA (50ng/µl)	0.5 µl	-
PCR Buffer with $MgCl_2$	1.0 µl	1X
dNTPs (2mM)	1.0 µl	0.2mM
F.P. (10pm/µl)	0.4 µl	0.4pm/µl
R.P. (10pm/µl)	0.4 µl	0.4pm/µl
Betaine (5M)	4.0	2M
DMSO	0.5	5%
Taq DNA Polymerase (5U/µl)	0.1 µl	0.5 U/ µl
MQ	5.2 µl	-
TOTAL	**10.0 µl**	

Appendix 3: Reference population to be studied

1. Tamang	Gaydan et al., 2007 and 2013	13. Udaya	Unpublished data (Shrestha et al., 2013)
2. Newar		14. Bajracharya	
3. Kathmandu		15. Shakya	
4. Tibet		16. Manandhar	
5. North China	Deng et al., 2005	17. S. and N. India	Thangraj et al., 2005
6. South China		18. Nepal_Kat & Mix	Wang et al., 2012
7. Han China		19. Burghats	Derenko et al., 2008
8. Tharu Eastern	Fornarino et al., 2009	20. Altain	
9. Tharu CI and CII		21. Uzbeck	
10. Hindu India		22. Korean	
11. Hindu Terai		23. Japanese	
12. AP Tribes		24. Mongolian	

Appendix 4: Mitochondrial D-loop region (rCRS Accession no. NC_012920)

TCCACCATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAGATTT
GGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCAT
GAATATTGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATCCACATCAAAACCCCC
TCCCCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCAC
CCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGT
ACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATGACCCCCCTCAGATAGGGGTCCCTTGACC
ACCATCCTCCGTGAAATCAATATCCCGCACAAGAGTGCTACTCTCCTCGCTCCGGGCCCATAACACTTGG
GGGTAGCTAAAGTGAACTGTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAGCCTAAATAGCCCACA
CGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTC
CATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCAC
CCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGG
CGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGC
ACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCAC
AGCACTTAAACACATCTCTGCCAAACCCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCAAACC
CCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAAC
ACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCACTTTTAACAGTCACCCCCCAACTAACAC
ATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAATACAACCCCCGCCCATCCTACCCAGCACACAC
ACACCGCTGCTAACCCCATACCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTAC
CTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCCATAAACAAATAGGTTTGGTCCT
AGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCC

Fig 18. Mitochondrial D-loop region. Green colour: primer M262, Blue colour: Primer M15976, Orange: Primer M16413, Gold colour: Actual D-loop region (1122 bp), Purple colour: repeated region in D-loop used for primer designing.

Appendix 5: PCR Amplified region of specific trinucleotide repeat loci in different SCAs, HD, DM1 and FXTAS

The forward and backward arrows indicate forward and reverse primers respectively. Associated trinucleotide repeats are marked by bold letters.

SCA1:

CAACATGGGCAGTCTGAGCCAGACGCCGGGACACAAGGCTGAG**CAGCAGCAGCAGCAGCAGC**
AGCAGCAGCAGCAGCAGCAT**CAG**CAT**CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA**
GCAGCAGCACCTCAGCAGGGCTCCGGGGCTCATCACCCCGGGGTCCCCCCCACCAGCCCAGCA
GAACCAGTACGTCCACATTTCCAGTTCTCCGCAGAACACCGGCCGCACC

Product size – 235 bp, Number of CAG repeats: 30

SCA2:

GGGCCCCTCACCATGTCGCTGAAGCCC**CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA**
GCAGCAACAGCAGCAGCAGCAGCAGCAGCAGCCGCCGCCCGCGGCTGCCAATGTCCGCA
AGCCCGGCGGCAGCGGCCTTCTAGCGTCGCCCGCCGCCGCGCCTTCGCCGTCCTCGTCCTCG
GTCTCCTCGTCCTCGGCCAC

Product size – 206 bp, Number of CAG repeats: 26

SCA3:

CCAGTGACTACTTTGATTCGTGAAACAATGTATTTTCCTTATGAATAGTTTTTCTCATGGTGTATT
TATTCTTTTAAGTTTTGTTTTTTAAATATACTTCACTTTTGAATGTTT**CAG**A**CAGCAG**CAAAAG**CA**
GCAA**CAGCAGCAGCAGCAGCAGCAGCAG**GGGGACCTATCAGGACAGAGTTCACATCCATGTG
AAAGGCCAGCCACCAGTTCAGGAGCACTTGGGAGTGATCTAGGTAAG

Product size – 241 bp, Number of CAG repeats: 12

SCA7:

AAGGAGCGGAAAGAATGTCGGAGCGGGCCGCGGATGACGTCAGGGGGGAGCCGCGCCGCGC
GGCGGCGGCGGCGGGCGGAG**CAG**CGGCCGCGGCCGCCCGG**CAGCAGCAGCAGCAGCAGCAG**
CAGCAGCAGCCGCCGCCTCCG**CAG**CCC**CAG**CGG**CAGCAG**CACCCGCCACCGCCGCCACGGCGC
ACACGGCCGGAGGACGGCGGGCCCGGCGCCGCCTCCACCTCGGCCGCCGCAATGGCGACGGTC
GGGGAGCGCAGGCCTCTGCCCAGTCCTGAAGTGATGCTGGGACAGTCGTG

Product size – 298 bp, Number of CAG repeats: 15

SCA8:

GGTCCTTCATGTTAGAAAACCTGGCTTTACTACTACTACTACTACTACTACTACTACTA**CTGCTGC**
TGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCATTTTTTAAAAATATATTATCTTATTT
TACTATTTGATGTTATAATTGTTATATATTTTTCCACACTTCCTCATACTGCTTATCTCTTACTTAAG
AATTTATGAATAAAGAATTGATTTTTCAATACATCCTTCCAAAAATTATCTGATGTTGAGTTAGTT
GCTCTCTCTTGTGCATTCTCAGTCCTCACAAGCCTTTCTCAAA

Product size – 309 bp, Number of CTG repeats: 15

SCA12:

TGCTGGGAAAGAGTCGTGGGGCTGCTGACGCGGTTGGGAGGAGCCTCGCCTTTAATGCACCAG
CCGCCTCCAGCCTCCTG**CAGCAGCAGCAGCAGCAGCAGCAGCAGCAG**CTGCGAGTGCGCGCGT
GTGGGTGTGAGGGTGAGTGCGCTGGC

Product size – 152 bp, Number of CAG repeats: 10

DRPLA:

CTCTTAGCCAACAGCAATGCCCATCCAGGCCCCTTCCCTCCCTCTACTGGGGCCCAGTCCACCGC
CCACCCACCAGTCTCAACACATCACCATCACCACCAGCAACAGCAA**CAGCAGCAGCAGCAGCAG**
CAGCAGCAGCAGCAGCAGCAGCAGCAGCATCACGGAAACTCTGGGCCCCCTCCTCCTGGAGCA
TTTCCCCACCCACTGGAGGGCGGTAGCTCCCACCACGCACACCCTTACGCCATGTCTCCCTCCCT
GGGGTCTCTGAGGCCCTACCCACCAGGGCCAGCACACCTGCCCCCACCTCACAGCCAGGTGTCC
TACAGCCAAGCAGGCCCCAATGGCCCTCCAGTCTCTTCCTCTTCCAACTCTTCCTCTTCCACTTCTC
AAGGGTCCTACCCATGTTCACACCCCTCCCC

Product size: 419 bp, Number of CAG repeats: 15

DM1:

AACGGGGCTCGAAGGGTCCTTGTAGCCGGGAATG**CTGCTGCTGCTGCTGCTGCTGCTGCT**
GCTGCTGCTGCTGCTGCTGCTGCTGCTGGGGGGATCACAGACCATTTCTTTCTTTCGGCCAG
GCTGAGGCCCTGACGTGGATGGGCAAACTGCAGGCCTGGGAAGGCTACAAGGACCCTTCGAGC
CCCGTTGATGGGCAAACTGCAGGCCTGGGAAG

Product size: 171 bp, Number of CTG repeats: 20

FXTAS:

GCTCAGCTCCGTTTCGGTTTCACTTCCGGTGGAGGGCCGCCTCTGAGCGGGCGGCGGGCCGAC
GGCGAGCGCGGGCGGCGGCGGTGACGGAGGCGCCGCTGCCAGGGGGCGTG**CGG**CAGCG**CGG**
CGGCGGCGGCGGCGGCGGCGGCGGCGGAGG**CGGCGGCGGCGGCGGCGGCGGCGG**CTG
GGCCTCGAGCGCCCGCAGCCCACCTCTCGGGGGCGGGCTCCCGGCGCTAGCAGGGCTGAAGAG
AAGATGGAGGAGCTGGTGGTGGAAGTGCGGG

Product Size – 278 bp, Number of CGG repeats: 20

HD:

ATGGCGACCCTGGAAAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTC**CAGCAGCAGCAGC**
AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCACCGCC
GCCGCCGCCGCCGCCGCCTCCTCAGCTTCCTCAGCCGCC

Product Size: 164 bp, CAG repeat number: 19

Appendix 6: Format of Consent Form

TRINUCLEOTIDE REPEAT LENGTH DISTRIBUTION AND MITOCHONDRIAL DNA HAPLOGROUP IN SUB-ETHNIC GROUP OF NEWAR POPULATION OF NEPAL

Study No: **Date:**

We would like your participation in this research as we need your blood sample to study the trinucleotide repeat length distribution and genetic linkage of Newar population. The purpose of this letter is to ensure your right to decide whether you want to participate in the research or not. You have the full right to ask questions if you are confused about the procedure. We will take 5ml of your blood. The whole procedure requires 10 minutes. The collected blood samples will be taken to India; Institute of Genomics and Integrative Biology, New Delhi where further research will be done. Confidentiality will be maintained regarding your identity. After the use of the blood samples in the current research, the result could be retained in Genetic Database for future use. While withdrawing blood you will not be harmed in any way.

Advantage:

There is no such definite personal advantage for being involved in this research. However, from this research the population database regarding trinucleotide repeat disorder and Genetic Data Base could be established.

Confidentiality:

This results for this research could be published, but your identity will not be revealed.

Agreement for Self-Participation:

You are participating for this research according to your will. You can withdraw from this research at any time without any hesitation. By signing below you agreeing that you have read, listened and your queries have been answered. Thus, you allowed for 5 ml blood collection entirely by your own wish.

Participant's Signature: …….
Participant's Name: …….
Age: …….
Address: ……
Date: ……
Contact no: ……

Only for illiterate:

I verify that I have read out and described all details to the above mentioned participant Mr/Mrs……. I am ensured that s/he has understood all requirements, s/he has given chance to ask questions and s/he has agreed to participate in this research. I verify that the finger print below is of the participants.

Field workers signature: …….
Field worker's name: …….
Date: ……. Participant's finger print